Privacy-Sensitive Audio Features for Conversational Speech Processing

THÈSE N^{O} 5234 (2011)

PRÉSENTÉE LE 6 DÉCEMBRE 2011

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR
LABORATOIRE DE L'IDIAP
PROGRAMME DOCTORAL EN GÉNIE ÉLECTRIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Sree Hari Krishnan PARTHASARATHI

acceptée sur proposition du jury:

Dr J.-M. Vesin, président du jury
Prof. H. Bourlard, Dr D. Gatica-Perez, directeurs de thèse
Prof. D. Ellis, rapporteur
Prof. S. King, rapporteur
Prof. J.-Ph. Thiran, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2011

Résumé

Le contexte de cette thèse est l'analyse des interactions sociales à partir de signaux audio de la vie de tous les jours. Pour atteindre cet objectif, nous souhaitons capturer des sons ambiants et conversationnels à l'aide d'un magnétophone portable. L'analyse de ces données peut s'appuyer sur une modélisation des tours de parole et leur répartition pour chaque locuteur. Toutefois, ces tâches se heurtent à des questions de respect de la vie privée dès lors que l'application finale se destine au grand public. Notamment, il est nécessaire d'obtenir le consentement explicite de chaque personne locuteur ne serait-ce que pour enregistrer ou stocker des données audio brutes.

Dans cette thèse, nous proposons d'utiliser des paramètres audio spécifiques qui, à la fois, respectent la vie privée en minimisant la quantité d'informations lexicales extraites du signal et permettent d'atteindre des performances de l'état de l'art pour certaines tâches de traitement de la parole. Plus précisément, nos principales contributions sont d'atteindre, à l'aide de ces paramètres audio, des performances dignes l'état de l'art pour les tâches de *détection de la parole* (c'est-à-dire la segmentation en zones de parole) et de *regroupement en locuteurs*. En complément, nous étudions le comportement de ces paramétres pour chacune de ces tâches dans des conditions variables : en intérieur (principalement) et en extérieur. Pour évaluer objectivement le concept de respect de la vie privée, nous proposons de faire transcrire le signal audio reconstruit par un système de reconnaissance automatique de la parole et par des individus, des taux de reconnaissance élevés signifiant que le respect de la vie privée est faible.

Concernant la tâche de détection de la parole fondée sur les paramétres respectueux de la vie privée, cette thèse étudie trois approches : (i) les méthodes d'extraction de caractéristiques simple et instantanée ; (ii) les méthodes d'extraction « d'excitation source » ; et (iii) les méthodes de « obfuscation »de l'audio. Ces approches ont été comparées aux coefficients PLP (*Prediction Linéaire et Perceptuelle*) dans de nombreuses conditions sur une grande base de données. En outre, nos études de reconnaissance automatique de la parole (phonèmes) sur la base de données TIMIT ont montré que les paramétres proposés conduisent à un respect de la vie privée plus élevé que les paramètres PLP.

Pour la tâche de regroupement en locuteurs, nous interprétons l'extraction des paramètres qui respectent la vie privée comme la recherche de l'ensemble optimal de paramètres qui, d'une part, maximise l'information mutuelle (IM) avec les locuteurs et, d'autre part, minimise l'IM avec les phonèmes. Le modèle ≪ source-filtre ≫ résulte naturellement de cette formulation. Ensuite, nous étudions deux approches différentes pour l'extraction des paramètres audio. Ces approches sont fondées sur le *résidu de la Prédiction Linéaire (PL)* et sur les *réseaux de neurones profonds*. À travers des expériences pour la tâche de regroupement en locuteurs faîtes à partir de données audio capturées par des microphones distants (isolés ou multiples), nous montrons que les paramètres proposés fournissent des performances proches des coefficients MFCC (*Mel Frequency Cepstral Coefficients*). D'autres résultats montrent que ces mêmes paramètres conduisent par ailleurs à des taux de reconnaissance moindres des transcriptions humaines en comparaison aux coefficients MFCC.

La dernière partie de la thèse étudie l'application de nos méthodes (détection de la parole et regroupement en locuteurs) à de l'audio enregistré en extérieur. Alors que notre étude du regroupement en locuteurs n'est qu'une étude préliminaire, nos résultats pour la détection de la parole amènent la conclusion selon laquelle les paramètres respectueux la vie privée fournissent des performances comparables aux paramètres PLP. Finalement, nous explorons la pertinence d'utiliser des modèles entraînés sur les conditions intérieures pour traiter de l'audio capturé en extérieur. Nous obtenons une large chute des performances que nous ne pouvons pas compenser, même en combinant les modèles appris pour l'intérieur.

Mots Clés : Audio respectueux de la vie privée, détection de la parole, regroupement de locuteurs, prediction linéaire, résidu, réseaux de neurones profonds, reconnaissance automatique de la parole, test d'intelligibilité.

Abstract

The work described in this thesis takes place in the context of capturing real-life audio for the analysis of spontaneous social interactions. Towards this goal, we wish to capture conversational and ambient sounds using portable audio recorders. Analysis of conversations can then proceed by modeling the speaker turns and durations produced by speaker diarization. However, a key factor against the ubiquitous capture of real-life audio is privacy. Particularly, recording and storing raw audio would breach the privacy of people whose consent has not been explicitly obtained.

In this thesis, we study audio features instead – for recording and storage – that can respect privacy by minimizing the amount of linguistic information, while achieving state-of-the-art performance in conversational speech processing tasks. Indeed, the main contributions of this thesis are the achievement of state-of-the-art performances in *speech / nonspeech detection* and *speaker diarization* tasks using such features, which we refer to, as privacy-sensitive. Besides this, we provide a comprehensive analysis of these features for the two tasks in a variety of conditions, such as indoor (predominantly) and outdoor audio. To objectively evaluate the notion of privacy, we propose the use of human and automatic speech recognition tests, with higher accuracy in either being interpreted as yielding lower privacy.

For the speech/nonspeech detection (SND) task, this thesis investigates three different approaches to privacy-sensitive features. These approaches are based on simple, instantaneous, feature extraction methods, excitation source information based methods, and feature obfuscation methods. These approaches are benchmarked against Perceptual Linear Prediction (PLP) features under many conditions on a large meeting dataset of nearly 450 hours. Additionally, automatic speech (phoneme) recognition studies on TIMIT showed that the proposed features yield low phoneme recognition accuracies, implying higher privacy.

For the speaker diarization task, we interpret the extraction of privacy-sensitive features as an objective that maximizes the mutual information (MI) with speakers while minimizing the MI with phonemes. The source-filter model arises naturally out of this formulation. We then investigate two different approaches for extracting excitation source based features, namely *Linear Prediction (LP) residual* and *deep neural networks*. Diarization experiments on the single and multiple distant microphone scenarios from the NIST rich text evaluation datasets show that these features yield a performance close to the Mel Frequency Cepstral coefficients (MFCC) features. Furthermore, listening tests support the proposed approaches in terms of yielding low intelligibility in comparison with MFCC features.

The last part of the thesis studies the application of our methods to SND and diarization in outdoor settings. While our diarization study was more preliminary in nature, our study on SND brings about the conclusion that privacy-sensitive features trained on outdoor audio yield performance comparable to that of PLP features trained on outdoor audio. Lastly, we explored the suitability of using SND models trained on indoor conditions for the outdoor audio. Such an acoustic mismatch caused a large drop in performance, which could not be compensated even by combining indoor models.

Keywords: Privacy-sensitive audio, speech/nonspeech detection, speaker diarization, linear prediction, residual, deep neural networks, speech recognition, listening tests.

Acknowledgements

It sounds almost clichéd to say that my PhD has been a long journey. But looking back over these four memorable years, pockmarked with unending deadlines, spent working at Idiap and housed in the most laid-back of places, time never felt a linear sensation – flying by at times, and standing still at others; a long journey it has been, but clichéd it never felt. Over this journey, from the first day of introductions to the day of my thesis defense, I am indebted to a number of people, both professionally and personally. A few, I am grateful to, even before it all began. It is both my duty and my pleasure to thank everybody involved in this personal journey.

Foremost, I am thankful to my thesis directors – Dr. Daniel Gatica-Perez and Prof. Hervé Bourlard. From Daniel, I have hopefully learnt not just research methodology, but also a professional outlook. I admire his approach to the creation of research problems. Gracias Daniel! With Hervé, I have always appreciated not only the independence that he gives his students, but also how he keenly critical he is. I admire his ability to abstract the essentials. Merci Hervé!

Over these years, as is natural in a PhD, I have had a number of interesting, technical discussions with many people – specifically, I am thankful to Mathew and Phil. I wish to acknowledge and appreciate Mathew's time and comments during the early part of my PhD. Thanks Mathew. With Phil, the discussions have been a bit more informal and a little less technical; but these discussions are now part of my system. "Mercios" Phil!

It was a real pleasure working with several researchers at Idiap. I value the interactions with Deepu, Fabio, Joel, and Sivaram – for their comments, scripts, and sometimes plain philosophy! I thank Gwénolé for correcting the abstract that I wrote in French (and later rewriting it almost completely!), and Harsha for proof-reading several parts of this thesis. Several colleagues from Idiap made it a happy workplace for me – Petr, Ferran, Sriram, Mike, Sarah, Ganga, Paco, Dayra, Serena, John, Dinesh, Radu, Kate, Gokul, Joan, Oya, Minh-Tri, Laurent, Lakshmi, Hui, Venky, Jagan, George, Zoltan, Rémi, Riwal, Thomas, Stefan, and Danil. There must surely exist a relationship between the Friday beer (and the group thereof) and a general reduction in the stress levels. Correlation may not amount to causation, but a "danke" to this group would not go amiss!

A big thanks to Idiap; specifically, I appreciate the great facilities. In this regard, I wish to thank the System Admin group for maintaining the computing grid, and also for addressing little problems such as providing laptops or printing posters. Merci Frank, Cedric, Bastien, Norbert, Tristan et Louis-Marie! I also wish to thank the secretariat and the finance group for helping me with several problems – be it preparing my residence permits or handling my apartments. Merci Sylvie, Nadine et Christophe! From a personal side, I am greatly indebted to Ed for his spontaneous help during a very difficult time in 2010. Thank you very much, Ed.

I would be terribly remiss if I did not acknowledge my appreciation to Prof. King, Prof. Ellis, Prof. Thiran, and Dr. Vesin for agreeing to be on my defense committee. I appreciate your time, and your positive comments motivate me! Thanks also to Swiss National Science Foundation for supporting this research through the MULTI and ICS projects. I also appreciate EPFL and its administrative group (Ms. Degott, Ms. Schneeberger, and Ms. Maier) for their support during (and especially the last part of) my PhD.

Two families have made me feel at home in Switzerland: my special thanks to Phil & Yoko and Hugues, Thomas, & their parents. I sincerely cherish your warmth and friendship. During the latter part of my stay here, I came to know Alex and I truly value our friendship. There are many locals/Valaisans with whom I have had interactions varying from "mild" to "intense". Some of these were truly game-changing moments: every time I think of them, I think of *Le petit prince*.

Finally, a big thank you to my family for their unconditional support. This thesis would not have been possible without you, and I dedicate this thesis to you. I also wish to thank my close friends (easy to forget the earlier era) for the many happy moments of gossip over the years. I invoke "privacy" to protect the gossip and the bearers!

Contents

List of Figures

List of Tables

Acronyms

Acronyms	**Definition**
AH	Features derived from non-initial maximum of the normalized autocorrelation, the number of autocorrelation peaks and the relative spectral entropy.
AMI	Augmented Multiparty Interaction.
AMIDA	Augmented Multiparty Interaction with Distance Access.
DCT	Discrete Cosine Transform.
DFT	Discrete Fourier Transform.
EZK	Features derived from energy, zero-crossing rate, and kurtosis.
FFT	Fast Fourier Transform.
GMM	Gaussian mixture model.
HMM	Hidden Markov model.
ICSI	International Computer Science Institute.
LP	Linear prediction.
LPR	Linear prediction residual.
LPRx	Features derived from linear prediction residual for a prediction order x.
MFCC	Mel frequency cepstral coefficients.
MFPLP	Mel frequency perceptual linear prediction.
MI	Mutual information.
MLP	Multilayer perceptron.
NIST	National Institute of Standards and Technology.
PCA	Principal Component Analysis.
PLP	Perceptual linear prediction.
RT	Rich Transcription.
SCD	Speaker change detection.
SEZK	Features derived from spectral flatness, energy, zero-crossing rate, and kurtosis.
SND	Speech/nonspeech detection.
SUS	Semantically unpredictable sentences.

Chapter 1

Introduction

Notwithstanding other forms of communication, speech is one of the most important channels of human interaction. On one hand, it is perhaps the most personal form of communication, providing a source of rich information, such as the message conveyed by the speaker, the identity of the speaker, and the emotion with which the words were said. On the other hand, it also allows us to infer higher levels of social interaction, such as the type of conversation (monologue, dyadic, or multiparty), the roles that people play, and the behavior that people exhibit.

In this context, there has recently been tremendous interest in capturing and analyzing spontaneous conversations using wearable devices. Besides increases in network bandwidth, this is due as much to the advances in the sensing technology as it is to the emergence of mobile phones and other devices as viable computing platforms. It is in this larger context that the work reported in this thesis takes place, with the aim of modeling face-to-face interaction patterns using audio collected with a portable recorder (Choudhury and Pentland, 2003; Gatica-Perez, 2006, 2009).

Although a number of interesting computational problems arise out of such a setting, for example, (a) computational complexity of algorithms for efficient utilization of battery; (b) unsupervised methods for handling large unlabeled data; and (c) analysis of social interactions using such audio data, this thesis limits itself to one of the principal issues in such ubiquitous data collection, namely privacy – of which more will be said subsequently. In this thesis, we mainly approach this problem from a speech processing viewpoint.

1.1 Overview of This Thesis

As discussed before, we wish to capture real-life audio using a portable recorder for the analysis of spontaneous social interactions. Analysis of conversations can then proceed, for instance, by modeling speaker turns using a speaker diarization system. However, without explicit consent from all people present, recording and storing spontaneous conversations leads to a breach in the privacy of people and it is usually unethical; sometimes this could even be illegal – but the exact legal terms are sometimes vague and are difficult to adjudicate, as it depends on many factors such as the laws of the country, sensitivity of the spoken information, etc. Among the earliest works that grappled with such a problem in the speech processing literature is Ellis and Lee (2004a,b), where the authors investigated the idea of using audio logs as a memory prosthetic. Much of the past work had assumed that audio (e.g. in dictation systems or personal recorders) was meant to be private, ie., produced and consumed at the individual level. This obviously changes in multiparty settings.

One way to address this issue of privacy is to store audio features instead of raw audio, such that neither intelligible speech nor linguistic content can be reconstructed (Ellis and Lee, 2004a; Wyatt *et al.*, 2007a). While such audio features may appear to be restrictive, there are different applications that with success use only the nonverbal cues in speech for the study of social behavior (Choudhury and Basu, 2004). Analysis of conversations using such features can then proceed by modeling the speaker turns and durations produced by speaker diarization. These features are referred to as privacy-sensitive (or privacy-preserving) features. The term “privacy-sensitive” can have different connotations in different areas of computing. Instead of coining a new term, this thesis decides to follow its use as originally proposed in the speech processing community by Wyatt *et al.* (2007a).

As an alternative to storing such audio features, one can implement a speech/nonspeech detection (SND) and a speaker diarization system on the portable device and store information based on the output. A caveat of this method though is that the set of possible tasks using such a high-level information is then limited by the output of the diarization system. For example, other sources of information, not including the verbal information, such as emotion, language, and the background acoustic scene information are inevitably lost. Another challenge concomitant with such a design choice is the computational limitations imposed by the portable device. Towards this end, sound sensing frameworks have begun to be proposed for the limited resources available on platforms like the Apple iPhone (Lu *et al.*, 2009).

This thesis considers the former approach to the problem, i.e., to capture and store audio features that respect privacy. Towards this, it studies audio features that can achieve state-of-the-art performance in *speech/nonspeech detection* and *speaker diarization* while minimizing the amount of linguistic information. A comprehensive analysis of various features for these tasks is performed in predominantly indoor conditions but also in outdoor conditions. This thesis then proposes methods to quantitatively assess privacy. In the next sections of this chapter, we will briefly discuss the motivations, objectives, contributions, and the organization of this thesis.

1.2 Motivations

Earlier works on privacy-sensitive features have studied applications of speech processing tasks such as scene analysis in indexing large, personal audio logs (Ellis and Lee, 2006) and multiparty conversation detection and modeling (Wyatt *et al.*, 2007a). One of the questions that arises naturally from these works for further investigation is the relative ability of different features to preserve privacy. Another motivation of this thesis is to systematically study such features for speech processing tasks and benchmark the features against standard features such as Mel Frequency Cepstral coefficients (MFCC).

The notion of privacy in audio is difficult to quantify and evaluate. Measures of usability of corrupted speech segments (Yantorno, 2000) could be interpreted as means to evaluate privacy, with high usability corresponding to low privacy. More recently, studies such as (Ellis and Lee, 2004a; Wyatt *et al.*, 2007a,b) indicate that the main privacy concern in audio is the reconstructibility of the linguistic information or of intelligible speech. For features such as MFCC, a potential issue is that both an intelligible speech signal and the linguistic content can be reconstructed.

In terms of spontaneous speech, meeting room recordings provide a challenging environment for many speech processing tasks (Garofolo *et al.*, 2004; Carletta *et al.*, 2006; Burger *et al.*, 2002; Janin *et al.*, 2003). State-of-the-art systems for such tasks use features based on spectral-shape based features. As examples, the ICSI meeting room diarization system (Wooters and Huijbregts, 2008) and the AMIDA 2009 meeting transcription system (Hain *et al.*, 2010) use features such as Mel

Frequency Cepstral coefficients (MFCC) and Mel Frequency Perceptual Linear Prediction (MFPLP) features (Dines *et al.*, 2006). While such features have been shown to be robust, as discussed earlier, they are not necessarily privacy-preserving.

In contrast to the audio recorded for conventional speech processing tasks, the audio from portable recorders present a different challenge; considering the portability of the recorders and the mobility that it provides the wearer, the features also need to be robust to changes in the ambient environment. Therefore, speech processing techniques, including feature extraction methods, need to be robust to a variety of conditions, such as indoor versus outdoor, matched versus mismatched, near-field versus far-field, and single versus multiple distant microphones.

1.3 Objectives

The ultimate aim of this work is to investigate privacy-sensitive features for tasks such as speech/nonspeech detection (SND), speaker change detection (SCD) and speaker diarization (SD) towards enabling the development of systems for conversation and acoustic scene analysis. This thesis has four specific objectives: (a) to investigate privacy-sensitive audio features for SND and benchmark them with respect to MFCC or PLP features; (b) to study state-of-the-art and privacy-sensitive features for speaker diarization exploring the tradeoff between task performance and privacy; (c) to propose an objective evaluation of the abstract notion of privacy; and (d) to assess the features in a variety of conditions – indoor and outdoor conditions.

1.4 Contributions

The main contributions of this thesis can be summarized as follows:

(1) State-of-the-art speech/nonspeech detection using privacy-sensitive features. This thesis investigates three approaches based on: (a) simple, instantaneous feature extraction methods; (b) excitation source information based methods; and (c) feature obfuscation methods such as local (within 130 ms) temporal averaging and randomization applied on excitation source information. To evaluate these approaches for SND, a multiparty conversational meeting data of nearly 450 hours is used. On this dataset, these features are benchmarked against standard spectral shape based features such as Mel Frequency Perceptual Linear Prediction (MFPLP). Fusion strategies combining excitation source with simple features show that comparable performance can be obtained in both close-talking and far-field microphone scenarios. As ways to evaluate privacy in audio, phoneme recognition studies are proposed, with higher recognition accuracy being interpreted as lower privacy. While excitation source features yield phoneme recognition accuracies in between the simple features and the MFPLP features, obfuscation methods applied on the excitation features yield low phoneme accuracies in conjunction with SND performance comparable to that of MFPLP features. This work was published by Parthasarathi *et al.* (2009a, 2010), and a detailed journal version was published by Parthasarathi *et al.* (2011b).

(2) Nearly state-of-the-art speaker diarization using privacy-sensitive features. We present an in-depth study of speaker diarization using privacy-sensitive features. Motivated by the source-filter model, Linear Prediction (LP) residual for speaker change detection is explored by Parthasarathi *et al.* (2009b). This was later extended to speaker diarization by (Parthasarathi *et al.*, 2011a). Issues such as prediction order and choice of representation of LP residual are studied. Additionally, the combination of LP residual with subband information and spectral slope is investigated. Subsequently, data-driven ways of estimating the excitation source are explored. In this regard, this thesis proposes a supervised framework using a deep neural network architecture for deriving privacy-sensitive audio features. These approaches are benchmarked against the traditional Mel Frequency Cepstral Coefficients (MFCC) features for speaker diarization in single and multiple distant microphone scenarios. Experiments on the NIST Rich Transcription Spring 2007 (RT07) evaluation dataset show that the proposed approaches yield diarization performance close to the MFCC features on the single distant microphone dataset. Human and automatic speech recognition tests are performed, showing that the proposed approaches to privacy-sensitive audio features yield much lower recognition accuracies compared to MFCC features. A detailed version of this work was submitted as a journal and is currently being revised (Parthasarathi *et al.*, 2011c).

(3) Mutual information framework. Extraction of privacy-sensitive features is interpreted as an objective that maximizes the Mutual Information (MI) with speakers while minimizing the MI with phonemes. The source-filter model arises naturally out of this formulation. Issues regarding estimation of MI, and an analysis with respect to different features were presented and published in (Parthasarathi *et al.*, 2011a).

(4) Outdoor conditions. The last contribution of this thesis is the validation of our approaches in outdoor conditions. While the diarization study in outdoor audio was more preliminary in nature, our study on SND confirms the conclusion that privacy-sensitive features trained on outdoor audio yield performance comparable to that of PLP features trained on outdoor audio. We also explored the suitability of using, and then combining SND models trained on indoor conditions for the outdoor audio. In this regard, we propose a maximum likelihood based unsupervised method to combine the SND models. Chapter 6 summarizes the results in outdoor conditions.

This thesis is based on the publications listed below.

Journals:

1. Parthasarathi et al. 2011b: Sree Hari Krishnan Parthasarathi, Daniel Gatica-Perez, Hervé Bourlard, and Mathew Magimai.-Doss, "Privacy-Sensitive Audio Features for Speech/Nonspeech Detection", *IEEE Transactions on Audio, Speech and Language Processing, 2011*.
2. Parthasarathi et al. 2012: Sree Hari Krishnan Parthasarathi, Hervé Bourlard, and Daniel Gatica-Perez, "Wordless Sounds: Robust Speaker Diarization using Privacy-Preserving Audio Representations", Under revision, *IEEE Transactions on Audio, Speech and Language Processing*.

Conferences:

1. Parthasarathi et al. 2011a: Sree Hari Krishnan Parthasarathi, Hervé Bourlard, and Daniel Gatica-Perez, "LP Residual Features for Robust, Privacy-Sensitive Speaker Diarization", *Proc. of Interspeech 2011*.
2. Parthasarathi et al. 2010: Sree Hari Krishnan Parthasarathi, Mathew Magimai.-Doss, Hervé Bourlard and Daniel Gatica-Perez, "Evaluating the Robustness of Privacy-Sensitive Audio Features for Speech Detection in Personal Audio Log Scenarios", *Proc. of ICASSP 2010*.
3. Parthasarathi et al. 2009b: Sree Hari Krishnan Parthasarathi, Mathew Magimai.-Doss, Daniel Gatica-Perez and Hervé Bourlard, "Speaker change detection with privacy-preserving

audio cues", *Proc. of ICMI-MLMI 2009*.

4. Parthasarathi et al. 2009a: Sree Hari Krishnan Parthasarathi, Mathew Magimai.-Doss, Hervé Bourlard and Daniel Gatica-Perez, "Investigating privacy-sensitive features for speech detection in multiparty conversations", *Proc. of Interspeech 2009*.
5. Parthasarathi et al. 2008: Sree Hari Krishnan Parthasarathi, Petr Motlicek, and Hynek Hermansky, "Exploiting contextual information for speech/non-speech detection ", *Proc. of TSD 2008, LNCS/LNAI series, Springer-Verlag, 2008*.

1.5 Organization

This thesis is organized as follows:

- Chapter 2 is an overview of the relevant literature on privacy. We start with a historical and legal perspective on privacy and then present a brief review of the existing literature on computational methods to preserve privacy in the ubiquitous computing and the distributed data mining communities. Relevant literature on privacy from speech processing is then surveyed.
- In Chapter 3, we present our study on speech/nonspeech detection using privacy-sensitive features. These features are benchmarked against the MFPLP features in terms of phoneme recognition on the TIMIT dataset and SND performance on the meeting room dataset.
- Before proceeding to speaker diarization, Chapter 4 presents an exploration of linear prediction (LP) residual features for the problem of speaker change detection on the HUB 4 dataset.
- In Chapter 5, we study the feature extraction as a process that seeks to maximize mutual information with speakers, while minimizing mutual information with phonemes. The source-filter model arises naturally out of this. We then view linear prediction (LP) residual in this light and then discuss the proposed deep neural network based features as another way to estimate the excitation source. Diarization experiments on single and multiple distant microphones are performed to compare proposed features with the standard MFCC features.
- Chapter 6 validates the proposed approaches to SND and speaker diarization on outdoor conditions. For SND, it proposes a maximum likelihood combination of privacy-sensitive detectors using the models trained on indoor conditions.
- Chapter 7 provides a final discussion of the work in this thesis and discusses future directions.

Chapter 2

Privacy: A Survey of Related Work

Among the many forms of privacy, this thesis is most concerned with information privacy. The intention of this chapter is by no means to provide a comprehensive review of information privacy and the methods to address it. Rather, it aims to portray the relevant literature in privacy-sensitive speech and audio processing within the larger fabric of legal and scientific ideas on privacy.

We start with some historical and legal interpretations of information privacy. Then we briefly look at guidelines for designing privacy enhancing techniques. Next, we review relevant literature on computational methods to preserve privacy. Specifically, we present interpretations and solutions drawn from the domains of ubiquitous computing and distributed data mining. Lastly, we turn to prior works from the field of speech processing. Here, we discuss the work on personal audio logs by Lee and Ellis. This is followed by the work done on conversation modeling using spontaneous speech by Choudhury, Wyatt, and others. Before concluding this chapter, we present a summary of recent work based on approaches from distributed data mining by Pathak and Raj.

2.1 What is Privacy and Who Cares About It?

Information privacy is not a recent phenomenon, and in one of the classic papers in the legal literature, Samuel Warren and Louis Brandeis (Warren and Brandeis, 1890), describe privacy as a "the right to be let alone". With advances in technology since then, the idea of privacy has evolved, but the basic notion of privacy now is as relevant as it was then. The following excerpt from Warren and Brandeis (1890) illustrates this point: “Instantaneous photographs and newspaper enterprise have invaded the sacred precincts of private and domestic life; and numerous mechanical devices threaten to make good the prediction that what is whispered in the closet shall be proclaimed from the house-tops.”

Subsequent to the publication of Warren and Brandeis (1890), there was considerable debate in the courtrooms over the question of whether the right of privacy existed at all (Prosser, 1960). Inquiry into the nature of interests that were being protected by privacy was still nascent.

2.1.1 Legal Viewpoints

Several decades later William Prosser expanded this idea based on torts. For this he synthesized various cases that emerged from Warren and Brandeis (1890) and listed four types of activities that are redressed under the rubric of privacy (Prosser, 1960):

- Intrusion upon the plaintiff’s seclusion or solitude, or into her private affairs.
- Public disclosure of embarrassing private facts about the plaintiff.
- Publicity which places the plaintiff in a false light in the public eye.
- Appropriation, for the defendants advantage, of the plaintiffs name or likeness.

This torts view of privacy, applicable between private parties, does not cover many of the subtle issues created by the capture of multimodal sensory data by organizations in ordinary social settings.

In a recent article for the Pennsylvania Law Review, Daniel Solove (Solove, 2006) addresses this by claiming that in order to understand privacy, one needs to identify how it is violated. He proposes a taxonomy of how an individual’s privacy can be violated. It starts with *information collection*, where various entities collect information about data subjects. Although Solove (2006) focuses mainly on surveillance and interrogation, this is relevant in the context of this thesis, as our focus is on the ubiquitous capture of audio. In the next stage of the taxonomy, *information processing*, some form of processing is done on the collected data. Briefly, this consists of five forms, namely aggregation, identification, insecurity, secondary use, and exclusion. The third is the *information dissemination* stage, which refers to outward propagation of the collected and processed information. The activities at this stage could range from breach of confidentiality to blackmail. The last stage is *invasion*. This stage does not involve information, as it mainly is related to what Solove refers to as intrusion and decisional interference. Out of the four stages, information collection and information processing are the stages for which privacy is incorporated commonly by design; and these form the focus of the rest of this chapter. It is in this context that the thesis is written.

2.1.2 Survey based Viewpoints

Such an analysis would be largely academic if people did not care sufficiently about privacy. In this regard, Alan Westin (Westin *et al.*, 2003) conducted about 30 surveys between 1978 and 2004, on a number of topics ranging from consumer privacy surveys to health information surveys and e-commerce. A summary of his conclusions is presented by Kumaraguru and Cranor (2005). In these surveys, based on their replies, Westin classified the people he surveyed into three categories. The first category are the *fundamentalists*, so called as they are "generally distrustful of organizations that ask for their personal information, worried about the accuracy of computerized information and additional uses made of it, and are in favor of new laws and regulatory actions to spell out privacy rights and provide enforceable remedies." The second class of people are the *pragmatists*, who "weigh the benefits to them of various consumer opportunities and services, protections of public safety or enforcement of personal morality against the degree of intrusiveness of personal information sought and the increase in government power involved." The last class are the *unconcerned*, who are "generally trustful of organizations collecting their personal information, comfortable with existing organizational procedures and uses, and ready to forego privacy claims to secure consumer-service benefits or public-order values."

Roughly 25% of the surveyed people were fundamentalists, and about 55% fall under the pragmatic category, while about 20% are unconcerned. This implies that a large majority of the people are concerned about privacy. This leads to a question of whether privacy can be factored into the design. To this end, we will look at Fair Information Principles (FIP). While the FIPs are not binding, they offer guidelines regarding data privacy.

2.2 Fair Information Principles

Langheinrich (Langheinrich, 2009b), in his article on privacy for ubiquitous computing, motivates the design of privacy-sensitive methods in ubiquitous computing from the Organization for Economic Cooperation and Development's (OECD) FIP. We take a similar approach and paraphrase the 8 principles; detailed information can be obtained from (OECD, 2011).

1. *Collection Limitation Principle:* There should be limits to the collection of personal data and any such data should be obtained with the knowledge or consent of the subject.
2. *Data Quality Principle:* Personal data should be relevant to the purposes for which they are to be used, and to the extent necessary for those purposes.
3. *Purpose Specification Principle:* The purposes for which personal data are collected should be specified before the time of data collection.
4. *Use Limitation Principle:* Personal data should not be disclosed, made available or otherwise used for purposes other than those specified in accordance with the Purpose Specification principle except with the consent of the data subject.
5. *Security Safeguards Principle:* Personal data should be protected by security safeguards against unauthorized access or disclosure of data.
6. *Openness Principle:* There should be a general policy of openness about developments, practices, and policies with respect to personal data.
7. *Individual Participation Principle:* An individual should have the right to obtain her data from the data controller within a reasonable time at a reasonable charge.
8. *Accountability Principle:* A data controller should be accountable for complying with the principle stated above.

Of course once can interpret the FIPs as one way to address the four stages of privacy violation of Solove (2006), and therefore provide a reasonable way to implement privacy. While many of these principles relate to policies, computational approaches have focused on the first, second, and fifth (security safeguards) principles. These form the topic of our next section.

2.3 Computational Methods to Preserve Privacy

Computational solutions to address privacy focus on the time of acquisition or on the time of processing. Due to the pervasive nature of data acquisition in ubiquitous computing, most prior work focus on acquisition; although some such as Hong and Landay (2004) do it at both stages. In contrast, in the field of distributed data mining, the focus has primarily been on the latter stage, through the use of secure multiparty computation. We will look at these approaches in some detail.

2.3.1 Ubiquitous Computing

Most prior work in ubiquitous computing has focused on data items such as location, Radio Frequency Identification (RFID), and smartspaces (Langheinrich, 2009b). Building smartspaces involves a fully privacy-sensitive infrastructure involving privacy beacons, privacy-sensitive database access, and privacy proxies. As this is beyond the scope of this thesis, we do not discuss smartspaces here. Details can be obtained from Hong and Landay (2004).

Location information is typically used by wearable devices (or mobile phones) or by web-based applications. In all these cases, the key questions are "Who? Where? When?". Knowing the answers to only one of these questions might not compromise the privacy so much. Computational methods to handle, exploit and either *obfuscate* the location (i.e., degrading the quality of location measurements) or *anonymize* (i.e., formally known as k-anonymity) the identity of the person have been proposed. A detailed survey of location privacy is provided in Krumm (2009).

Among the various technologies in ubiquitous computing, RFIDs are the most prominent in terms of privacy issues (Langheinrich, 2009a). The typical scenario is that of an RFID tag attached to a consumer item that is being tagged. The tag is typically passive, meaning that it receives power from a reader. The literature on RFIDs is more involved, as they tend to reveal location traces as well as other personal data (depending on the item). Privacy of the data is ensured not only by an encrypted communication but also by an authentication mechanism, the latter of which requires a shared secret – also known as pass key – to be selected and exchanged between the tag and the reader. Two important challenges posed by passive RFIDs are the limited computational resources of the tag and the selection/exchange of pass key. A common response to this is to use a powerful proxy device that would locally manage one's tagged items. By incorporating an RFID reader into a device such as a mobile phone, the device could act as a proxy for all interactions: blocking or allowing access to the tagged items (Langheinrich, 2009a).

2.3.2 Distributed Data Mining

Privacy in distributed data mining investigates secure information processing over distributed data. A scenario considered commonly is that of two independent agents, with one agent requiring access to some private data, with the other securing all the access to it. The basis of all methods in this case is the classical Secure Multiparty Computation (SMC) protocol introduced by (Yao, 1986). This is usually illustrated through the Millionaires' problem: suppose two people want to find out who is worth more, but neither want to reveal their worth. (Yao, 1986) presented a solution to this problem for two parties, and it has since been extended to the multiparty case (Goldreich *et al.*, 1987).

Formally in a multiparty case, the goal is to compute a global function, without any party knowing anything except their own input and the global value. For instance, if the global function is a computation of the sum of the values, then every party should not be able to know more than the sum of the inputs of the other parties. If all parties can be assumed to be honest, a simple solution is: the first person adds a random value to her input and passes it to the second, who adds his input to this and so on until it cycles back to the first person, who subtracts the random value to yield a "secure sum". This algorithm is popular in data mining and has been applied for many machine learning algorithms such as large margin training for SVMs (Vaidya *et al.*, 2008) and clustering (Lin *et al.*, 2005), among others.

2.4 Privacy in Speech and Audio Processing

In the previous section, while surveying privacy from two fields of research, we categorized the literature into three privacy enhancing techniques, namely anonymization, obfuscation, and cryptographic methods. In the field of speech processing, apart from this thesis, there are three lines of work on privacy. In this section, we will give a brief outline of these works.

2.4.1 Scene Analysis in Long Duration Personal Audio Logs

Although the idea of using wearable devices to capture real-life audio has been explored before (Clarkson *et al.*, 1998), to our knowledge the earliest works on privacy in audio emerged from Ellis and Lee (2004b). In this work, inspired by Vannevar Bush's vision of memex, the authors used audio to record the daily experience. Features derived from a long-term average (of the order of a minute) of standard short-term spectral features are used in conjunction with the Bayesian Information Criterion (BIC) to segment locations.

In their subsequent publications (Ellis and Lee, 2004a, 2006), they observed that while speech is perhaps the richest and the more interesting content, storing raw recordings inevitably leads to privacy concerns. (Ellis and Lee, 2004a) then goes on to define a notion of privacy-sensitivity based on the linguistic message, "Given a reliable method for speech segment identification, we propose to have our systems default behavior be to scramble such segments to render the words unintelligible." To this end, in Lee and Ellis (2006) short-term features based on autocorrelation were proposed for robust speech/nonspeech segmentation. These features are meant to be used for detecting speech segments and making them unintelligible before storage.

2.4.2 Conversation Analysis

In the field of automatic conversation analysis, (Basu, 2002) is probably among the earliest work on audio features. Here a dyadic conversation analysis – conversation detection, segmentation, and classification – is performed using a series of detectors (e.g., voicing detectors) based on audio features derived from short-term autocorrelation and relative spectral entropy. These audio features were studied for robustness to noise, robustness to microphone distance, and robustness to environment. This study was done using raw audio, and no explicit privacy constraints were included.

Choudhury (2004) is, to our knowledge, the first work on sensor-based data-driven methods for modeling face-to-face interactions. Here, a "sociometer" was equipped to collect the following information for each individual: (a) identity of the people in an interaction, as obtained by using an infrared transceiver sending/receiving a unique ID; (b) speech information sampled at 8 KHz; and (c) motion information captured using an accelerometer. Although raw audio was recorded and stored, it was agreed with the participants that only speech features proposed by Basu (2002) would be extracted; the underlying social network structure was then learned from these features (Choudhury *et al.*, 2003; Choudhury and Pentland, 2003).

The next generation of sociometers (Olguin-Olguin, 2007) did not store raw audio, and instead, stored energy values from four frequency bands. These frequency bands were, 85 to 222 Hz, 222 to 583 Hz, 583 to 1527 Hz, and 1527 to 4000 Hz. Offline analysis of these values was done to extract higher-level cues such as, speech activity, consistency of speech, and the amount of speaking time per minute. These and other face-to-face interaction cues were studied by Olguin-Olguin and Pentland (2010) in three separate case studies of "organizational design and engineering".

In an attempt at explicitly addressing privacy, Wyatt *et al.* (2007a,b) reinterpreted (Basu, 2002)'s features – derived from short-term autocorrelation and relative spectral entropy – for privacy-sensitive, multiparty conversation detection and modeling. These works lend support to the notion expressed earlier in (Ellis and Lee, 2004a), of privacy in audio as a constraint on the features such that neither intelligible speech nor linguistic content be reconstructible (Wyatt *et al.*, 2007a). Furthermore, these features were recorded for a social network analysis on 4,400 hours of privacy-sensitive audio data captured from real world interactions, in completely unconstrained and natural conditions, between 24 subjects over a period of 9 months.

2.4.3 Distributed Data Mining based Methods

In the field of distributed data mining, one assumes the existence of a computational agent who acts independently of the distributed data. The idea behind this is expressed thus (Pathak and Raj, 2011): "Alice has some data for which she wishes to do some inference, while Bob possesses a model for this; The constraint is that while Alice wants the data to remain private, Bob wants the model parameters to remain private." Drawing upon methods from distributed data mining, (Pathak *et al.*, 2010) investigate methods for privacy of the the independent agent as well as the distributed data. For example (Lin *et al.*, 2005), shows how expectation maximization (EM) for training GMMs (Lin *et al.*, 2005) is done.

In particular, (Pathak and Raj, 2011; Pathak *et al.*, 2010) study these methods for problems such as keyword spotting, speaker identification, speaker verification, and speech recognition. They have extended (Lin *et al.*, 2005) to Hidden Markov Models (HMM).

2.5 Context of This Thesis

The following are some aspects which connect this thesis to the work presented in this chapter.

- Although earlier works such as Ellis and Lee (2004a); Wyatt *et al.* (2007a) indicate that the linguistic message is the privacy-sensitive information, a question that arises from these works for further investigation is the assessment of the relative privacy of different features. In this context, this thesis proposes speech recognition and intelligibility studies to benchmark privacy in features against MFCC and raw audio. In addition, since most face-to-face interaction tasks exploit SND and speaker diarization (Gatica-Perez, 2009), we also use MFCC/PLP as a yardstick to assess how the proposed features, as well as those by Wyatt *et al.* (2007a); Ellis and Lee (2004a) compare.

- Linear prediction has a long history in speech processing, with applications not only in speech coding and synthesis, but even in speech recognition. The source-filter model assumed by linear prediction provides a natural framework for exploring privacy concerns. As an extension, data-driven models assuming the source-filter model (not including linear prediction) could also be explored, as we have done with deep neural networks. This approach contrasts with those taken by Ellis and Lee (2004a) and Wyatt *et al.* (2007a). Additionally, one of the issues we investigate in this thesis is the choice of representation of the linear prediction residual for tasks such as SND and diarization.
- We explore privacy-sensitive methods based on obfuscation, thereby providing a link to the privacy preserving approaches in ubiquitous computing (Krumm, 2009) and those in speech processing (Ellis and Lee, 2004a).
- As a formal result of our study on SND, we formulated a model for privacy preserving methods using mutual information. To our knowledge, this framework has not been studied before.

Chapter 3

Speech/nonspeech Detection

The overall objective of this thesis is to study privacy-sensitive features for speech/nonspeech detection and speaker diarization. This chapter restricts itself to the first aspect; i.e., privacy-sensitive features for speech/nonspeech detection (SND). Towards this, we present a comprehensive study of privacy-sensitive features for SND in indoor meetings.

This work investigates three different approaches based on: (a) simple, instantaneous feature extraction methods; (b) excitation source information based methods; and (c) feature obfuscation methods such as local temporal averaging and randomization applied on excitation source information. To evaluate these approaches for SND, we use multiparty conversational meeting data of nearly 450 hours. On this dataset, we evaluate these features and benchmark them against standard spectral shape based features such as Mel Frequency Perceptual Linear Prediction (MFPLP). Fusion strategies combining excitation source with simple features show that a performance comparable to MFPLP can be obtained in both close-talking and far-field microphone scenarios. As one way to objectively evaluate the notion of privacy, we conduct phoneme recognition studies on TIMIT. While excitation source features yield phoneme recognition accuracies in between the simple features and the MFPLP features, obfuscation methods applied on the excitation features yield low phoneme accuracies in conjunction with SND performance comparable to that of MFPLP features. This work was originally published in Parthasarathi *et al.* (2009a, 2010, 2011b).

The rest of the chapter is organized as follows. Section 3.1 motivates our features. Section 3.2 provides an overview of our approach. The dataset definition and the annotations, including the dataset protocol involving various experimental setups are provided in Section 3.3. Section 3.4 discusses the implementation and the notational details of the SND system, comprising the features, the classifier, and the combination techniques. Parameter selection experiments are discussed in Section 3.5. We discuss the SND performance and revisit the privacy-sensitive aspects of the features in Sections 3.6 and 3.7. Finally, we draw some conclusions in Section 3.8.

3.1 Motivation and Preliminary Work

In this section, we motivate our three approaches, namely, (a) *simple, instantaneous features*, (b) *linear prediction residual*, and (c) *feature obfuscation methods* such as local (within 130 ms) temporal averaging and randomization.

3.1.1 Preliminary Work on Simple Features

Motivated by Wyatt *et al.* (2007a), we began our study into privacy-sensitive features for speech detection (Parthasarathi *et al.*, 2009a), by exploring four different, classical short-term features obtained by temporal processing of the audio signal (i.e., without estimating the spectrum). These features are energy, zero crossing rate, spectral flatness, and kurtosis. In addition to these four features, we also systematically studied the features proposed earlier by (Wyatt *et al.*, 2007a). A brief summary of the features and the conclusions of the study is now presented.

1. Short-term energy (E): Studies have shown that short-term energy has been one of the most important features for speech detection (Atal and Rabiner, 1976; Wrigley *et al.*, 2005). Furthermore, studies have shown that long-term information in energy can be exploited for speech detection (Parthasarathi *et al.*, 2008).

2. Short-term zero crossing rate (Z): The zero crossing rate at a frame-level has been a popular feature for voiced/unvoiced/nonspeech classification (Atal and Rabiner, 1976; Wrigley *et al.*, 2005).

3. Short-term spectral flatness measure (S): The short-term spectrum of the nonspeech signal such as wideband noise can be expected to be flatter than the short-term spectrum of the speech signal. Thus a measure of spectral flatness can be useful for SND. By normalizing the

spectrum, and viewing it as a probability mass function, entropy can be used as a measure of the flatness of spectrum. Linear prediction analysis can also be used to derive an efficient flatness measure of speech without explicitly estimating the spectrum (Makhoul, 1975). The flatness measure is derived as the ratio of the energy in the model error (residual) to the energy in the original signal. We investigate the measure of spectral flatness obtained using the latter approach.

4. Short-term kurtosis (K): Kurtosis is derived from the fourth order moment of a distribution and it measures its "peakedness". Speech samples have been shown to have a flatter distribution than noise samples, and kurtosis has been shown to be useful in this regard (Wrigley *et al.*, 2005).

5. Features proposed in (Basu, 2002; Wyatt *et al.*, 2007a) for privacy-sensitive speech detection (AH) are the non-initial maximum of the normalized autocorrelation, the number of autocorrelation peaks and the relative spectral entropy.

Benchmarking the two sets of *simple features*, proposed by (Parthasarathi *et al.*, 2009a) and (Wyatt *et al.*, 2007a), revealed that the performance of these privacy-sensitive features with explicit temporal modeling is comparable to the standard spectral features such as MFPLP, that do not have the privacy constraint. Our subsequent study focusing on the robustness of these features, (Parthasarathi *et al.*, 2010), however found that there could be a small gap in performance between the privacy-sensitive and the non privacy constrained features in *mismatched* conditions.

3.1.2 Motivation for Linear Prediction Residual

It is generally known that at least two or three formants are required to synthesize intelligible speech or to reconstruct the linguistic information (Donovan, 1996). Our approach to preserving privacy is based on adaptively filtering out information about these spectral peaks. This approach is motivated by the source-filter model (Fant, 1960).

Linear prediction (LP) analysis of speech (Makhoul, 1975) assumes the source-filter model and it estimates three components: (a) an all-pole model; (b) a residual; and (c) a gain;. The vocal tract response is modeled by the all-pole model, with the model capacity being determined by the prediction order (p). The LP residual, obtained by inverse filtering the speech signal with the all pole model, can be considered to be privacy-preserving.

Depending on the prediction order, the LP residual contains mostly information about the excitation source of the speakers (Prasanna *et al.*, 2006). It has been shown that humans can recognize speakers by listening to the LP residual signal (Feustel *et al.*, 1989). Previous works have exploited this. For example, the LP residual has been used as a complimentary feature for speaker recognition in (Thevenaz and Hugli, 1995), while (Prasanna *et al.*, 2006) exploits speaker information in the LP residual at segmental levels (10 - 30 ms) using an autoassociative neural network.

Another property of LP residual is that it has been shown to be relatively robust to additive noise (Murty *et al.*, 2007). The Hilbert envelope of the LP residual is processed in (Murty *et al.*, 2007) using covariance analysis and the periodicity property of this signal was then used in a voice activity detection task.

The importance of long temporal context ($\approx$250 ms) for spectral-shape based features such as MFCC is well known for ASR (Hermansky and Morgan, 1994). This has also been exploited for SND in (Dines *et al.*, 2006). In this chapter, we investigate whether information at such temporal scales exist in LP residual.

Our work extends these previous works in several ways. Unlike (Thevenaz and Hugli, 1995) we use LP residual independent of the all-pole model parameters. Secondly, in contrast to (Murty *et al.*, 2007) and (Prasanna *et al.*, 2006) we investigate and then exploit long temporal context in LP residual. A systematic investigation of the LP residual for various prediction orders is conducted for SND. The robustness of the LP residual in farfield microphone data is then evaluated. To the best of our knowledge, the present chapter (Parthasarathi *et al.*, 2011b) is the first work that exploits LP residual in a privacy-sensitive SND scenario.

3.1.3 Obfuscation Methods

As discussed in Section 2.3, obfuscation methods have been used previously in other aspects of privacy in sensor data research (Krumm, 2009). In this chapter, we apply these techniques, in particular, averaging and randomization, on MFPLP and LP residual features. We provide more details in Section 3.4.1.

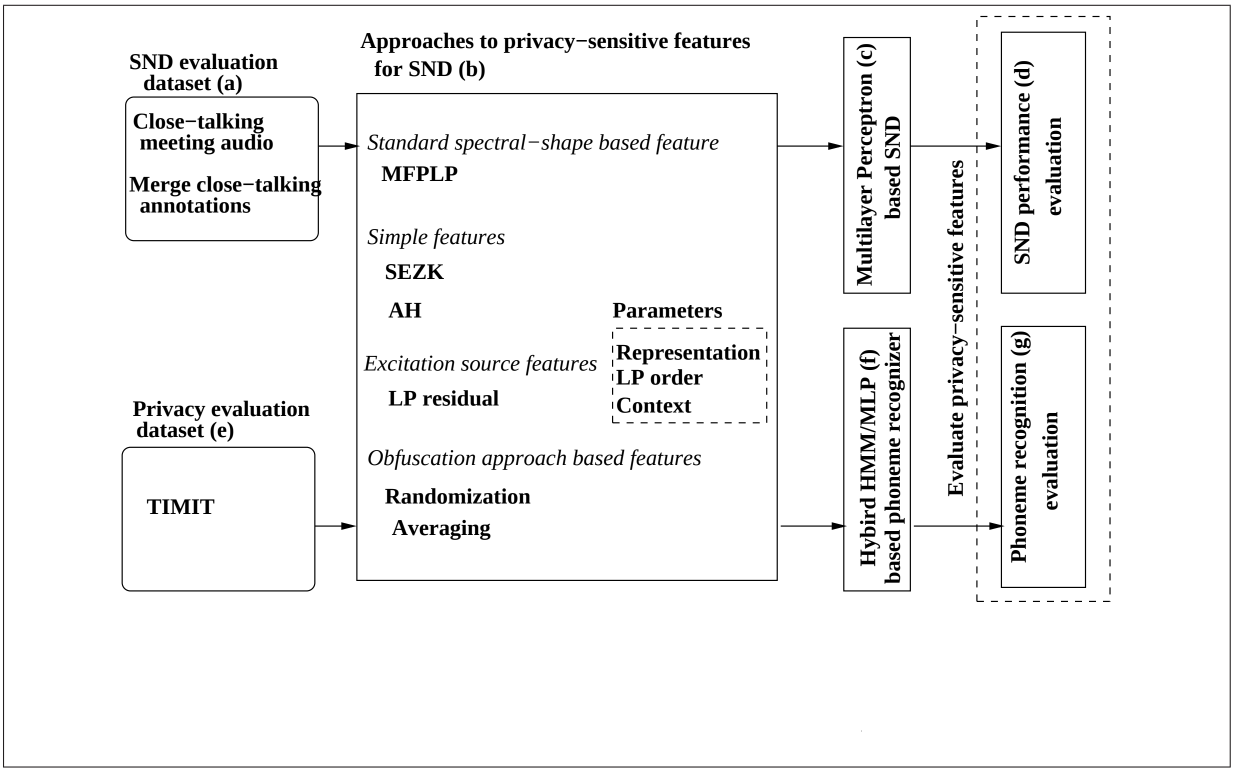
SND evaluation dataset (a)
Close−talking meeting audio
Merge close−talking annotations
Privacy evaluation dataset (e)
TIMIT
Approaches to privacy−sensitive features for SND (b)
Standard spectral−shape based feature
MFPLP
Simple features
SEZK
AH
Parameters
Representation
LP order
Context
Excitation source features
LP residual
Obfuscation approach based features
Randomization
Averaging
Multilayer Perceptron (c) based SND
Hybird HMM/MLP (f) based phoneme recognizer
Evaluate privacy−sensitive features
SND performance (d) evaluation
Phoneme recognition (g) evaluation

Figure 3.1: Block diagram of our approach. A detailed discussion of the figure is provided in Section 3.2.

3.2 Our Approach to This Chapter

Figure 3.1 illustrates our approach to this chapter using a block diagram. These blocks are described below.

(a): Evaluating privacy-sensitive features for speech detection entails a comparison of SND performance as well as an evaluation of linguistic privacy. To evaluate SND we construct the scenario using multiparty meeting data, namely the NIST (Garofolo *et al.*, 2004), AMI (Carletta *et al.*, 2006), and ICSI (Janin *et al.*, 2003) databases. Section 3.3 discusses the SND datasets in more detail.

(b): Privacy-sensitive and the standard spectral features (MFPLP) are derived from these datasets. Some issues with LP residual are the choice of parameters, namely, its representation, the LP order, and the temporal context. Section 3.5 describes parameter selection experiments with these features, their combinations, and the notations in detail.

(c,d): A separate multilayer perceptron (MLP) classifier is trained for each feature set for the speech/nonspeech classification task, similar to (Dines *et al.*, 2006). This allows us to compare the privacy-sensitive features with the reference MFPLP features, by way of eliminating the effects of the classifier. MLP classifier is also useful in studying the effect of temporal context. Section 3.4.2 provides more details on the MLP classifier, while Section 3.6 presents the SND results.

(e,f,g): The notion of linguistic privacy is quantified using phoneme recognition studies on the TIMIT dataset. These experiments are performed with the hybrid HMM/MLP system (Bourlard and Morgan, 1994). The trained MLP classifiers used for SND are different from the ones used in the hybrid HMM/MLP system. The phoneme recognition results with these features are provided in Section 3.7.

3.3 SND Dataset

An issue in comparing the features is a lack of standard datasets, due to privacy concerns. For this study, we used the scenario that was constructed in (Parthasarathi *et al.*, 2009a). We likened the audio collected by subjects wearing portable audio recorders to a meeting room scenario captured using close-talking microphones. In contrast to the traditional meeting room applications where, given the close-talking microphone signal, the interest generally lies in the speech segments of the wearer ((Dines *et al.*, 2006), (Wrigley *et al.*, 2005)), in conversation analysis, speech segments that are spoken by the other speakers are also of interest. As a consequence of this, crosstalk segments in the meeting room tasks are now considered as speech segments.

3.3.1 Dataset and Annotations

The dataset and annotations were used from our setup in (Parthasarathi *et al.*, 2009a). The audio data consists of individual close-talking microphone recordings from meetings. Groundtruths are then derived by merging the speech-activity annotations for the individual microphones, that are closer than a fixed time interval of 100 ms. Since manual annotations are not consistent ((Fiscus *et al.*, 2006)), forced-alignment was used to derive the annotations for the individual microphones. More details on the forced-alignment procedure used to derive the annotations can be found in (Dines *et al.*, 2006).

A figure illustrating this merging process for a close-talking microphone recording of a meeting room speech segment is shown Figure 3.2. Each speaker's SND annotation for that meeting segment is done with respect to whether the speaker spoke or not during that segment. During this

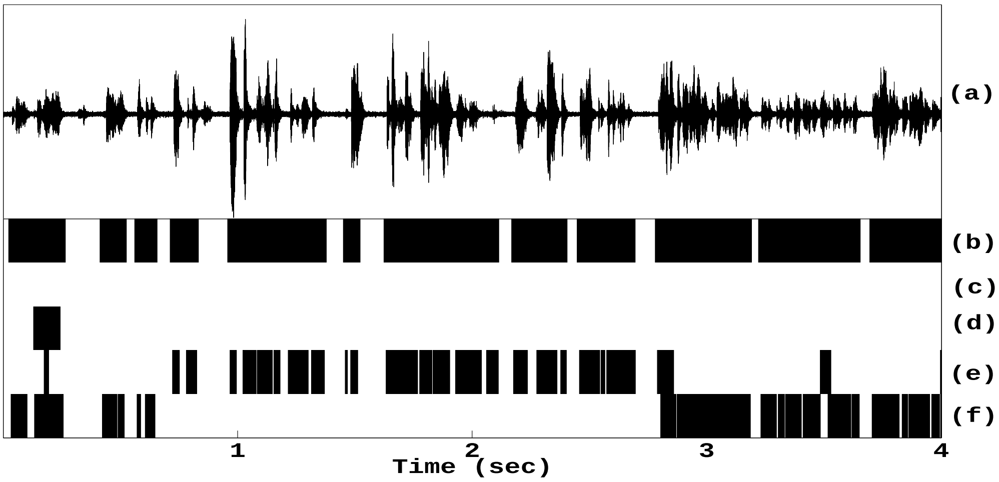

Figure 3.2: A close-talking microphone recording of a meeting segment with the speech/nonspeech annotations on the four close-talking microphone channels and their merged annotation. Dark regions indicate speech segments and light regions indicate nonspeech segments. (a) close-talking microphone recording (b) merged annotation using the annotations from all the channels (c) Speaker 1's annotation - appears to be silent in this segment (d) Speaker 2's annotation for the same meeting segment with respect to her microphone recording (e) Speaker 3's annotation for the same meeting segment with respect to her microphone recording - in this case, the signal in Fig. 3.2(a) was used to produce this annotation (f) Speaker 4's annotation for the same meeting segment with respect to her microphone recording.

meeting segment, speakers 1 and 2 (Figure 3.2(c), (d)) appear to be mostly silent. The annotation corresponding to the wearer of this microphone is shown in Figure 3.2(e). The merged groundtruth using the process discussed above is shown in Figure 3.2(b).

The close-talking microphone recordings, sampled at 16kHz, were obtained from NIST (Garofolo *et al.*, 2004), AMI (Carletta *et al.*, 2006), and ICSI (Janin *et al.*, 2003) meeting room data. The total data adds up to 100 hours of meeting speech spanning 120 meetings. The actual amount of individual close-talking recordings adds up to nearly 450 hours with NIST, AMI and ICSI contributing 52, 50 and 350 hours respectively. The training data from NIST, AMI and ICSI amounted to 9, 15 and 48 hours respectively. Using the groundtruth defined above, the overall ratio of nonspeech to speech was around 1:4.2. The amount of near-field speech is considerably less than the amount of far-field speech, with overall ratio of nonspeech: near-field speech:far-field speech being 1.4: 1: 4.8.

3.3.2 Dataset Protocol

Using the dataset described earlier, we construct matched, mismatched, and cross-validation conditions. The notations for these conditions are described in Table 3.1. Numbers inside brackets denote the number of hours and the numbers outside denote the notation for that particular dataset.

For a training dataset x and a test dataset y from the table, we use the notation $\{N \,\text{or}\, A \,\text{or}\, I\}\{N \,\text{or}\, A \,\text{or}\, I\}xy$, where N, A, and I correspond to NIST, AMI, and ICSI datasets respectively. The 3 matched setups on NIST, ICSI, and AMI used in (Parthasarathi *et al.*, 2009a) are NN14, AA25, and II36 respectively. Similarly the 6 mismatched setups used in (Parthasarathi *et al.*, 2010) are NA15, NI16, AN24, AI26, IN34, and IA35 respectively. The cross-validation setups are AI23 and IA32.

Table 3.1. Train and test datasets for matched, mismatched and cross-validation experiments. Numbers in the brackets denote the number of hours and the numbers outside denote the notation for that dataset.

Features	NIST	AMI	ICSI
Train	1 (9)	2 (15)	3 (48)
Test	4 (52)	5 (50)	6 (350)

3.4 SND System

As part of the experimental setup, all SND systems have been constrained to have access to audio from one channel only. This section discusses the implementation and the notational details of the features, followed by the MLP classifier. Combinations of classifiers and features are discussed next.

3.4.1 Features

All the features are extracted by pre-emphasizing the signal and then using a 25 ms analysis window with a 10 ms shift.

Simple features

The first set of simple features are spectral flatness (S), energy (E), zero-crossing rate (Z), and kurtosis (K) (Parthasarathi *et al.*, 2009a). In our implementation of short-term spectral flatness, it is derived as the ratio of the energy of the LP model error (residual) to the energy of the original signal (Makhoul, 1975). The energy feature is implemented as short-term log-energy of the signal, while kurtosis feature is implemented as the short-term signal kurtosis. We use *SEZK* and *EZK* to denote the set of all four features and the set of three features respectively.

The features proposed in (Wyatt *et al.*, 2007a) and (Basu, 2002) are the non-initial maximum of the normalized autocorrelation, the number of autocorrelation peaks, and the relative spectral entropy. The relative spectral entropy feature is implemented as the Kullback-Leibler divergence between the normalized power spectrum of the current frame and a normalized average of the power spectra of the previous 500 frames (Basu, 2002). Let *AH* denote this feature set.

Based on our previous works ((Parthasarathi *et al.*, 2009a), (Parthasarathi *et al.*, 2010)), the temporal context is fixed at 51 frames and the features are augmented with their first and second derivatives. The dimensionalities of *SEZK*, *EZK*, and *AH* for each frame are 12, 9, and 9 respectively.

Linear Prediction Residual Based Features

We now look at some issues in using LP residual as features.

(a) Choice of representation of the LP residual: The representations of the residual studied are: a real-cepstrum representation ((Thevenaz and Hugli, 1995)) with a fixed number of 12 coefficients along with c_0 and a MFPLP representation with 12 coefficients along with c_0. The MFPLP representation is computed using HTK (Young *et al.*, 2000). These features are augmented with delta and acceleration coefficients. Feature selection experiments investigating the choice of representation are presented in detail in Section 3.5. In either representation, with delta and acceleration coefficients, the dimensionality of the LP residual features for each frame is 39. Delta and acceleration coefficients of LP residual yielded a small gain in performance on the cross-validation data.

(b) LP order: We study LP residual by varying the prediction orders from 2 to 20. The choice of the LP order presents a tradeoff between privacy and SND performance.

(c) Temporal context: The efficacy of temporal context for LP residual with respect to the SND task is studied by varying the temporal support from no-context (1 frame) to 101 frames (with 50 frames for both left and right context).

Temporal Obfuscation Approach

The two obfuscation methods studied are:

(a) Local temporal randomization: Feature vectors within a block of size ($N = 1, 5, 9, 13$) are shuffled. A uniform pseudo-random number generator was used to shuffle the frames in the block. It can be noted that a randomization of N frames could result in two successive frames being separated by $2 \cdot (N-1)$ frames (equivalently $2 \cdot (N-1) \cdot 10$ ms). We chose block sizes up to 13 frames because results in (Pinto *et al.*, 2011) indicate that phonetic information in the speech signal up to 230 ms (12 frames each way) can be exploited for phoneme recognition.

(b) Local temporal averaging: Feature vectors within block of size ($N = 1, 5, 9, 13$) are averaged. These methods are applied to MFPLP and LP residual based features.

Table 3.2. Number of input and hidden units for each MLP.

Features	Input	Hidden
Simple features	$51 \times$ dim of feature	200
LP residual features	$\{1, 31, 51, 101\} \times$ dim of feature	50
LP residual with simple features	$51 \times$ dim of feature	100
MFPLP features	$31 \times$ dim of feature	50

Spectral-Shape based Features (MFPLP)

The 12 MFPLP coefficients along with c_0 are computed using HTK. In addition, log-energy and signal kurtosis are extracted. Delta and acceleration coefficients are then appended. In (Dines *et al.*, 2006), these features were augmented with a set of cross-channel based features. Since we use each microphone channel independently, we drop the cross-channel based features, while we retain all the other features. We use the notation *MFPLP* to denote this feature set. The total dimensionality of this feature set for each frame is 45.

3.4.2 MLP based SND Classifier

A separate MLP classifier was trained on each feature set for speech/nonspeech targets based on the groundtruth definition described in Section 3.3. The minimization of cross-entropy was used as the training criterion. All the features are normalized to zero-mean and unit variance at the input of the MLP using the global means and variances estimated on the training data. The number of hidden and input units in the MLP classifier trained for simple features and MFPLP features were identified by model selection in our previous studies Parthasarathi *et al.* (2009a). For the LP residual features these experiments were conducted on the cross-validation set. These results are summarized in Table 3.2.

3.4.3 Classifier and Feature Combinations

Classifier combination techniques (Kittler *et al.*, 1998) typically combine either the decisions made by the individual classifiers or assign a weight to each classifier's evidence to exploit complementary information. These weights can be either estimated statically or dynamically. In our experiments, we explored one static and one dynamic classifier combination technique: (a) Averaging the weights - static weighting (b) Inverse entropy weighting - dynamic weighting. However, from our study on mismatched conditions for SND (Parthasarathi *et al.*, 2010), and from the experiments performed for (Parthasarathi *et al.*, 2011b), it was observed that averaging the weights performed consistently better. For the sake of clarity, we only present our studies on averaging the weights.

Feature-level combinations are also studied to investigate the possibility of exploiting the correlation between features. To this end, feature-level combinations of the LP residual based features and the simple features are investigated.

3.4.4 SND Evaluation Measure

For SND evaluation, we use the Area under the Receiver Operating Characteristics (AROC) curve as a metric to evaluate speech detection, as in (Parthasarathi *et al.*, 2009a; Wrigley *et al.*, 2005; Parthasarathi *et al.*, 2010) . The Receiver Operating Characteristics (ROC) curve is plotted by varying the detection-threshold on the posterior probability estimates provided by the SND MLP. A value of 50% for the AROC indicates a random performance and value of 100% indicates a perfect classification. Furthermore, this measure was selected so that the evaluation measure is not biased towards a prior distribution of speech and nonspeech.

On the other hand, one of the issues in using measures like AROC is the difficulty in establishing significance intervals. One approach to mitigate this is to perform a Chi-squared test on the count of the meetings over which one of a pair of two systems is better. While this is a weak test, in cases where the significance is “obvious”, it can confirm it. We do not perform this test in this thesis, and it is a limitation of this work.

3.4.5 Notation

For the discussions that follow in this chapter, the notation for the feature sets, the MLP based SND systems, and the combinations are summarized in Table 3.8. In the table, the notation $F(x)$ stands for an MLP based system trained for a feature (or a feature set) *x*. For example, $F(E_1)$ is an MLP based system trained on energy with no context but with delta and acceleration coefficients. Similarly, $F(EZK_{51})$ is an MLP based system trained on energy (E), zero-crossing rate (Z), and kurtosis (K) with 51 frame context and with delta and acceleration coefficients. To explicitly indicate feature-level combinations of simple features with LP residual based features, we use the notation: $F(x, y)$. For example, $F(LPR8_{51}, EZK_{51})$ denotes a feature-level combination of the individual features ***LPR8*** and ***EZK*** using 51 frame context.

We use $C(x, y)$ to denote a system obtained by combining the output of individual MLP systems based on features x and y using classifier combination. For example, the system $C(LPR8_{51}, EZK_{51})$ performs a classifier combination of the individual systems $F(LPR8_{51})$ and $F(EZK_{51})$.

3.5 Parameter Selection for LP Residual Features

We now conduct studies on the parametrization of LP residual: (a) choice of representation of LP residual; (b) LP prediction order; and (c) effect of temporal context on LP residual. These studies were performed on the cross-validation set, namely, AI23 and IA32. The optimal hyperparameters are fixed for later studies in Section 3.6 and 3.7.

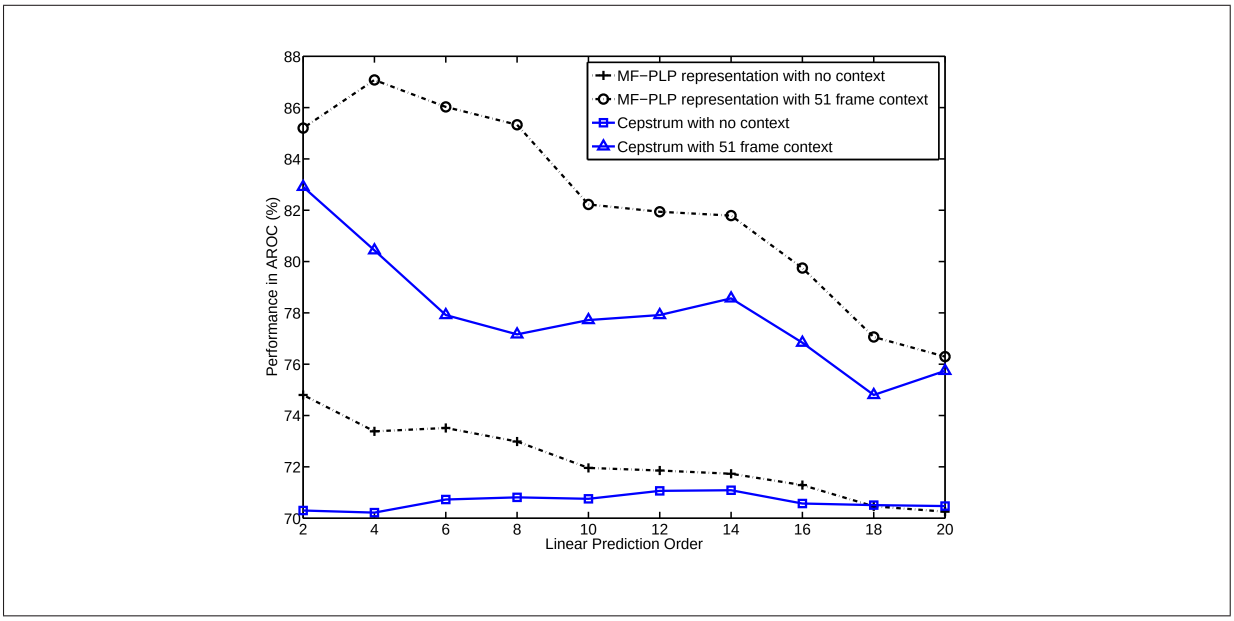

Figure 3.3: Choice of representation of linear prediction residual on AI23 dataset. The two representations of the residual studied are cepstrum and MFPLP. The x-axis is the linear prediction order and y-axis is the SND performance in area under the receiver operating curve (AROC). This figure also compares the two representations with two different temporal contexts - no context and 51 frame context.

3.5.1 Representation of LP Residual

We study the 2 choices of representations of LP residual discussed in Section 3.4.1: MFPLP and cepstral representation. Figure 3.3 shows the comparison between the 2 representations with two different temporal contexts - no context and 51 frames context on the AI23 dataset. It can be observed that MFPLP representation yields a better performance with both temporal contexts. This trend was observed on IA32 dataset as well.

3.5.2 Prediction Order

We now focus on the MFPLP representation in Figure 3.3 and investigate the choice of LP order. As the prediction order increases, the all pole model approximates the envelope of the short-time power spectrum better. Consequently, we see a drop in the performance for SND as the prediction order is increased. We note that the LP residual contains both modeling and excitation errors. As the LP order increases beyond 10, the contribution of the error in the residual signal is mainly due to the excitation error component.

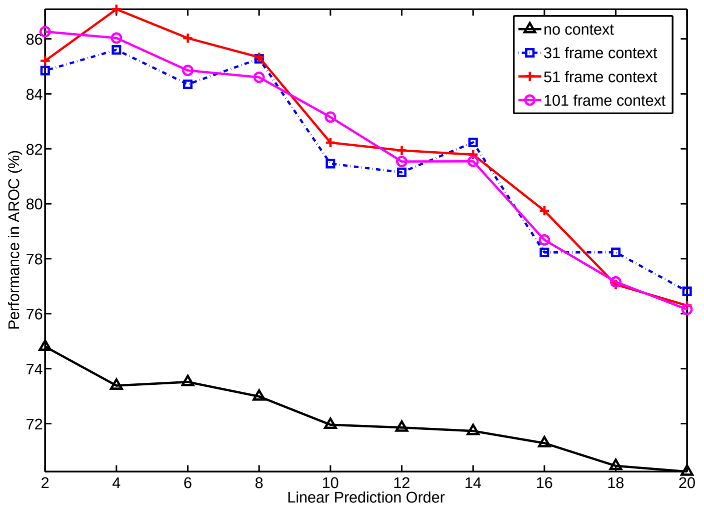

Figure 3.4: Effect of temporal context on the MFPLP representation of linear prediction residual on AI23 dataset. This plot shows four different temporal contexts - no context, 31 frame context, 51 frame context, and 101 frame context. The x-axis is the linear prediction order and y-axis is the SND performance in area under the receiver operating curve (AROC).

The vocal tract system is typically characterized by up to five resonances in the 0 to 4 kHz range. An LP order in the range of 8 to 14 can model between 2 to 5 formants. Revisiting the performance versus privacy tradeoff, an LP order of 8 seems appropriate for the SND task with a privacy constraint, since the first two formants are important for synthesizing an intelligible speech signal (Donovan, 1996).

3.5.3 Temporal Context

Figure 3.4 compares the SND performance when the temporal context of the LP residual features is increased. This plot shows four different temporal contexts - no context, 31 frame context, 51 frame context, and 101 frame context. A substantial gain in performance can be observed when the temporal context is increased from 1 frame to 31 frames. In general, there is a small gain for most LP orders when the context is increased from 31 frames to 51 frames. An increase in context from 51 frames to 101 frames does not yield any gain. For $F(MFPLP_{31}, EK)$, on the other hand, we observed that the performance saturates at around 31 frames. This observation is consistent with studies in (Dines *et al.*, 2006). These trends were observed on IA32 dataset as well.

3.5.4 Selected Parameters

To conclude this section, we fix the values of the following hyperparameters: (a) LP residual representation is MFPLP, (b) LP order is 8, and (c) Temporal context is 51 frames.

3.6 SND Performance Results

This section presents the results for simple features and excitation source features on matched and mismatched conditions. Further analyzes are performed on close-talking and far-field microphone recording scenarios. Feature-level and classifier-level combinations are also investigated. The next section 3.7 discusses phoneme recognition results to quantify privacy. As a means to enforce stricter privacy on excitation source features in terms of phoneme recognition rates, we then discuss the obfuscation methods.

Table 3.3. Performance of features (in percentage of area under ROC) with a context of 51 frames, in matched and mismatched conditions. The second column lists the overall performance of each system. N, A, and I refer to NIST, AMI, and ICSI datasets. $A \to B$ refers to the system being trained on a dataset A and being tested on a dataset B. The table is grouped into blocks of privacy-sensitive and non privacy-sensitive features. For each column and for each block, the best performance is highlighted in bold. The dataset protocol, mentioned in Section 3.3.2, is also mentioned for the respective columns in the table.

Features	All	N NN14	A AA25	I II36	N→A NA15	N→I NI16	A→N AN24	A→I AI26	I→N IN34	I→A IA35
		Matched conditions			Mismatched conditions					
Privacy-sensitive features										
$F(EZK_{51})$	80.5	84.1	90.8	82.0	82.0	75.5	86.0	80.3	82.5	86.7
$F(SEZK_{51})$	79.4	84.0	91.5	81.5	79.7	71.5	86.7	80.6	83.6	87.2
$F(AH_{51})$	81.3	83.3	90.3	85.7	86.0	75.7	85.3	78.9	83.6	88.1
$C(EZK_{51}, AH_{51})$	**83.4**	86.0	91.5	86.2	87.2	78.1	87.5	82.7	85.0	89.1
$F(LPR8_{51})$	84.8	83.0	90.9	89.0	84.5	79.6	83.4	85.3	83.3	87.8
$F(LPR8_{51}, SEZK_{51})$	84.7	86.7	91.3	88.9	85.6	79.0	86.7	84.1	**86.0**	87.6
$F(LPR8_{51}, EZK_{51})$	85.2	86.1	91.1	89.5	84.2	80.2	86.6	84.5	84.9	88.4
$C(LPR8_{51}, SEZK_{51})$	85.0	86.8	**92.1**	88.3	86.9	79.2	87.0	85.3	85.5	89.3
$C(LPR8_{51}, EZK_{51})$	85.4	86.7	91.8	88.6	87.2	81.1	86.9	84.6	85.1	89.1
$C(LPR8_{51}, AH_{51})$	**86.3**	86.1	91.8	**89.8**	88.4	**82.0**	86.4	**86.1**	85.1	89.8
$C(LPR8_{51}, EZK_{51}, AH_{51})$	86.0	**87.5**	92.0	88.9	**88.7**	81.8	**87.8**	85.4	**86.0**	**90.0**
Non privacy-sensitive features										
$F(MFPLP_{31}, \overline{DA})$	81.8	83.0	**91.6**	89.8	82.9	65.6	85.5	**86.8**	84.3	89.7
$F(MFPLP_{31})$	83.6	83.4	91.4	**90.7**	85.3	71.5	85.1	86.4	85.0	**90.2**
$F(MFPLP_{31}, EK, \overline{DA})$	83.0	**84.6**	91.1	87.9	84.9	73.5	**86.5**	84.8	84.3	88.4
$F(MFPLP_{31}, EK)$	**85.0**	84.5	**91.6**	89.9	**87.4**	**77.2**	86.1	86.3	**85.3**	90.0

The results are reported in Table 3.3 for NIST, AMI, and ICSI meeting data. In the discussion that follows, N, A, and I refer to NIST, AMI and ICSI datasets. $A \rightarrow B$ refers to the system being trained on a dataset A and being tested on a dataset B. The dataset protocol, mentioned in Section 3.3.2, is also mentioned for the respective columns in the table.

The second column lists the overall performance of each system. We observe that the combination of simple features yields benefit over individual systems, with exception of the addition of spectral flatness to $F(EZK_{51})$ (Parthasarathi *et al.*, 2010). LP residual based systems yield better performance than simple features. However, combinations of simple features with LP residual yield substantial gain in performance. For example, the best performing simple feature based system, $C(EZK_{51}, AH_{51})$, yields 83.4% while the best performing system with LP residual, $C(LPR8_{51}, AH_{51})$, yields 86.3%. Furthermore, we see that this system gives comparable or better performance than $F(MFPLP_{31}, EK)$ (85.0%). We note that the addition of delta and acceleration features, in addition to energy and kurtosis, yields gains to $F(MFPLP_{31}, \overline{DA})$.

We now further analyze the features in both matched and mismatched conditions.

3.6.1 Analysis on Matched Conditions

From Table 3.3, it can be seen that the performance of the LP residual based SND system with a context of 51 frames, denoted by $F(LPR8_{51})$ is slightly less than $F(EZK_{51})$, $F(SEZK_{51})$, $F(AH_{51})$ and $F(MFPLP_{31}, \overline{DA})$ for the NIST dataset. On the AMI dataset, all the features are comparable. Whereas, for the ICSI dataset, the LP residual is significantly better (at least 3%) than EZK, $SEZK$ and AH and it is comparable to $F(MFPLP_{31}, \overline{DA})$.

Next, we consider the feature combination studies. Table 3.3 shows that on matched conditions, $F(LPR8_{51}, SEZK_{51})$ and $F(LPR8_{51}, EZK_{51})$ yield superior performance in comparison with $F(EZK_{51})$, $F(SEZK_{51})$, and $F(AH_{51})$. These systems are comparable with the systems based on $MFPLP$ on all the three datasets.

Between the classifier combination schemes, our experiments revealed that the average weighted combination scheme is superior to the inverse entropy weighted scheme. For the sake of clarity, we do not present it in the table. However, the difference between the average weighted classifier combination scheme and the feature combination method is marginal (less than 1%).

Combining either *AH* or *EZK* features with residual based features through classifier combination scheme yields similar results. In matched conditions combining both *AH* and *EZK* with the residual based features through classifier combination methods does not yield consistent improvements over combinations with just one of the feature sets.

We now analyze the performance of $MFPLP$ features. It can be noted that the addition of delta and acceleration coefficients or energy and kurtosis to $F(MFPLP_{31}, \overline{DA})$ does not increase the performance significantly. In matched conditions it appears that simple spectral based system $F(MFPLP_{31}, \overline{DA})$, is sufficient for state-of-the-art performance.

Finally, the best performance for the privacy-sensitive features on the NIST, AMI, and ICSI datasets are 87.5%, 92.1%, and 89.8% respectively. The best performances achieved by the non privacy-sensitive features on the same datasets are 84.6%, 91.6%, and 90.1% respectively. We see that both sets of features are comparable on matched conditions.

Table 3.4. SND performance analysis (in percentage of area under ROC) in matched and matched conditions (AA25, II36). The table is grouped into blocks of privacy-sensitive and non privacy-sensitive features. For each column and for each block, the best performance is highlighted in bold. The dataset protocol, mentioned in Section 3.3.2, is also mentioned for the respective columns in the table. CT and FF refer to close-talking and far-field microphone scenarios.

	AMI AA25		ICSI II36		A→I AI26		ICSI dataset
	Matched conditions				Mismatched conditions		Summary
Features	CT	FF	CT	FF	CT	FF	FF
Privacy-sensitive features							
$F(E_1)$	89.7	74.6	92.4	68.4	92.0	68.4	68.4
$F(K_1)$	89.6	75.7	92.7	68.2	90.8	68.5	68.4
$F(Z_1)$	64.3	61.9	54.9	51.0	57.0	51.9	51.5
$F(E_{51})$	94.1	86.0	96.0	73.5	92.8	72.5	73.0
$F(K_{51})$	94.3	86.9	95.5	74.4	92.2	72.5	73.5
$F(Z_{51})$	88.9	80.0	81.8	67.1	82.0	61.0	64.1
$F(EZK_1)$	90.3	77.8	92.0	70.4	89.7	71.0	70.7
$F(EZK_{51})$	95.3	90.2	95.0	79.4	93.4	78.0	78.7
$F(AH_{51})$	94.9	89.8	96.1	84.1	91.6	77.2	80.7
$F(LPR8_{51})$	95.1	90.4	96.1	87.8	94.5	83.7	85.8
$C(LPR8_{51}, AH_{51})$	96.0	91.4	**97.5**	**88.5**	**96.1**	**84.3**	86.4
$C(LPR8_{51}, EZK_{51}, AH_{51})$	**96.2**	**91.7**	97.4	87.5	**96.1**	83.6	85.6
Non privacy-sensitive features							
$F(MFPLP_{31}, \overline{DA})$	**95.1**	91.3	**95.4**	85.3	**95.4**	**85.3**	85.3
$F(MFPLP_{31}, EK)$	94.8	**91.4**	**95.4**	**88.9**	93.2	85.1	87.0

3.6.2 Analysis on Mismatched Conditions

For the mismatched conditions, it can be seen that the LP residual based SND system is generally better than $F(EZK_{51})$ and $F(SEZK_{51})$. The comparison with $F(AH_{51})$ and $F(MFPLP_{31}, \overline{DA})$ is more mixed for $F(LPR8_{51})$.

Combining LP residual with ***SEZK*** at feature-level yields a small, if any, gain in performance. Comparison between $F(LPR8_{51}, EZK_{51})$ and $F(LPR8_{51}, SEZK)$ systems yields mixed results. Similar to matched conditions, $F(LPR8_{51}, SEZK_{51})$ and $F(LPR8_{51}, EZK_{51})$ yield superior performance in comparison with $F(EZK_{51})$, $F(SEZK_{51})$, and $F(AH_{51})$.

In contrast to feature combination methods, the classifier combination methods typically yield a bigger and a more consistent gain. Again, the average weighted combination scheme is superior to the inverse entropy weighted scheme. We do not include inverse entropy scheme in the table since the trends were similar to the average weighted combination scheme. Furthermore, from Table 3.3, we observe that, similar to feature combinations, $C(LPR8_{51}, EZK_{51})$ yields mixed results in comparison with $C(LPR8_{51}, SEZK_{51})$. This shows that the addition of the spectral flatness measure does not add significant complementary information to the classifier. Combining either ***AH*** or ***EZK*** features with residual based features through classifier combination schemes yields similar results.

Unlike the matched conditions, combining ***AH*** with residual based features appears to be better than combining ***EZK*** with residual based features through classifier combination methods. For example, on the A→I column, $C(LPR8_{51}, AH_{51})$ yields a performance of 86.1% while $C(LPR8_{51}, EZK_{51})$ yields a performance of 84.6%. Furthermore, unlike the matched conditions, combining all the three privacy-sensitive systems through classifier combination methods, yields, in general, a more consistent gain in performance than combining just two of them.

Regarding the performance of the spectral-shape based features, it can be noted that the addition of delta and acceleration coefficients to MFPLP coefficients yields a more consistent gain than in the matched condition case. Adding energy and kurtosis, also in general, yields improvements. The addition of delta and acceleration in conjunction with energy and kurtosis also yields a consistent gain in performance.

3.6.3 Analysis on Close-talking and Far-field Microphones

To gain better understanding, we further analyze the features with respect to close-talking and far-field microphones. In general, we expect the close-talking data in the matched conditions to be the easiest, while far-field data in the mismatched conditions to be the hardest. This was done not only to evaluate the privacy-sensitive features in these conditions, but also to investigate if the performance gains due to temporal context and due to feature/classifier combinations are consistent under all conditions.

To perform this analysis, the two-class groundtruth for SND on the test set was split into a three-class groundtruth: close-talking speech, far-field speech, and nonspeech. Close-talking groundtruths corresponding to the close-talking microphones were used for generating the three-class groundtruths. ROC curves are plotted for $\{\text{close-talking speech}, \text{nonspeech}\}$ and $\{\text{far-field speech}, \text{nonspeech}\}$, and the area under the ROC (AROC) is computed.

Analysis on Matched Conditions

The results are listed in Table 3.4. It can be observed from the table that for all single features such as energy, zero-crossing, and kurtosis, the increase in performance due to an increase in context is more significant in the far-field case than the close-talking case. For energy, for instance, due to the increase in temporal context, the gain in performance is nearly 12% in the far-field case on the AMI dataset, whereas, for the close-talking case, the gain due to increase in context is less than 5%. Similar trends can also be observed for ICSI dataset for the single features. Furthermore, even when no context is used, combinations of single features yield a bigger gain for the far-field case than the close-talking case.

Next, we analyze the performance of systems based on spectral-shape based features. As we had noticed in previous experiments (Parthasarathi *et al.*, 2009a), in comparison with AMI meetings, ICSI meetings were recorded in a larger meeting room with speakers being farther apart. This results in the signal-to-noise ratio (SNR) of the speech signal of a speaker who is farther from a close-talking microphone to be lower. We had hypothesized that spectral features such as $MFPLP$ handle this case more effectively. This is indeed observed to be true when we compare $F(MFPLP_{31}, \overline{DA})$ (85.3%) with $F(EZK_{510})$ (79.4%) and $F(E_{510})$ (73.5%) using the far-field scoring, for ICSI dataset in the matched conditions.

While performing at similar performance levels to $F(EZK_{510})$ on AMI near and far-field evaluations, we also observed that $F(LPR8_{51})$ performs significantly better when evaluated on ICSI far-field dataset. Furthermore, we see that the LP residual has complementary information compared to $F(AH_{51})$. Also, LP residual performs similar to $MFPLP$ features on the ICSI dataset using the far-field scoring in matched conditions. Lastly, in matched conditions, for the far-field scoring, combining both *AH* and *EZK* with residual based features through classifier combination methods does not yield consistent improvements over combinations with just *AH*.

Analysis on Mismatched Conditions

Table 3.4 also presents the results for the far-field and the close-talking cases in mismatched conditions. From the table, we observe a similar trend for a single feature such as energy, wherein there is an increase in performance due to an increase in context for the far-field scenario. But it is interesting to note that when there is no context, the performance of $F(EZK_1)$ is similar or slightly worse than $F(E_1)$ and $F(K_1)$ for the close-talking case, while it is better than all the three single features for the far-field scenario. On the other hand, when there is a temporal support of 51 frames, $F(EZK_{51})$ is consistently better than energy based and kurtosis based systems for both the close-talking and the far-field case.

We observe that $F(LPR8_{51})$, while performing at slightly better levels than $F(EZK_{510})$ on close-talking evaluations, performs significantly better when evaluated on far-field data. This along with the observations in the matched conditions case, strongly suggests that excitation based features $F(LPR8_{51})$ are robust not only with respect to distance, but also robust with respect to mismatched ambient conditions. This result is supported by robustness studies on LP residual such as (Murty *et al.*, 2007).

In the A→ I mismatched scenario we have chosen for this table, the spectral-shape based features yield the best performances in both close-talking and far-field scenarios. We have omitted the other mismatched conditions since the trends were similar. LP residual features in combination with simple features, show performance comparable to $MFPLP$ features in other far-field scenarios.

3.7 Revisiting Privacy

So far we have investigated simple features and LP residual based features. Before we investigate the temporal obfuscation approach, we briefly revisit privacy. To the best of our knowledge, quantitatively benchmarking audio features for privacy has not been studied before in the literature. Some possible ways to benchmark linguistic privacy in audio features could be: (a) human speech recognition rates of the synthesized speech from the privacy-sensitive features (b) subjective assessments of the privacy-sensitivity of features by human subjects (c) automatic speech recognition rates using the privacy-sensitive features. Since synthesizing speech using simple features is not trivial, we prefer ASR studies for quantifying privacy. ASR accuracies are generally reported in the literature using phoneme recognition rates or word recognition rates. The latter is more complex for assessing privacy due to the differences in vocabulary sizes, dictionaries, and language models.

3.7.1 Dataset for Phoneme Recognition

Phoneme recognition studies were performed on TIMIT database (4.3 hours), sampled at 16kHz. Experiments were conducted excluding the 'sa' dialect sentences. The training data consists of 3000 utterances from 375 speakers, cross-validation data consists of 696 utterances from 87 speakers, and the test data set consists of 1344 utterances from 168 speakers. The hand-labeled dataset using 61 labels is mapped to the standard set of 39 phonemes (Lee and Hon, 1989).

3.7.2 Phoneme Recognition System

Features are mean/variance normalized across the training data set. A three layered MLP is used to estimate the phoneme posterior probabilities. MLP consists of 1000 hidden units, and 39 output units with softmax nonlinearity, representing the phoneme classes. The input layer uses a temporal context of 9 frames on the features generated at a frame rate of 100 Hz, with delta and acceleration coefficients. The MLP is trained using standard back propagation algorithm by minimizing the cross entropy error criterion. The phoneme recognition experiments are performed using the hybrid HMM/MLP system reported in Bourlard and Morgan (1994). The phoneme sequence is decoded using the Viterbi algorithm, where each phoneme is represented by a left-to-right, 3-state HMM, enforcing a minimum duration of 30 ms. The emission likelihood in each of the three states is the

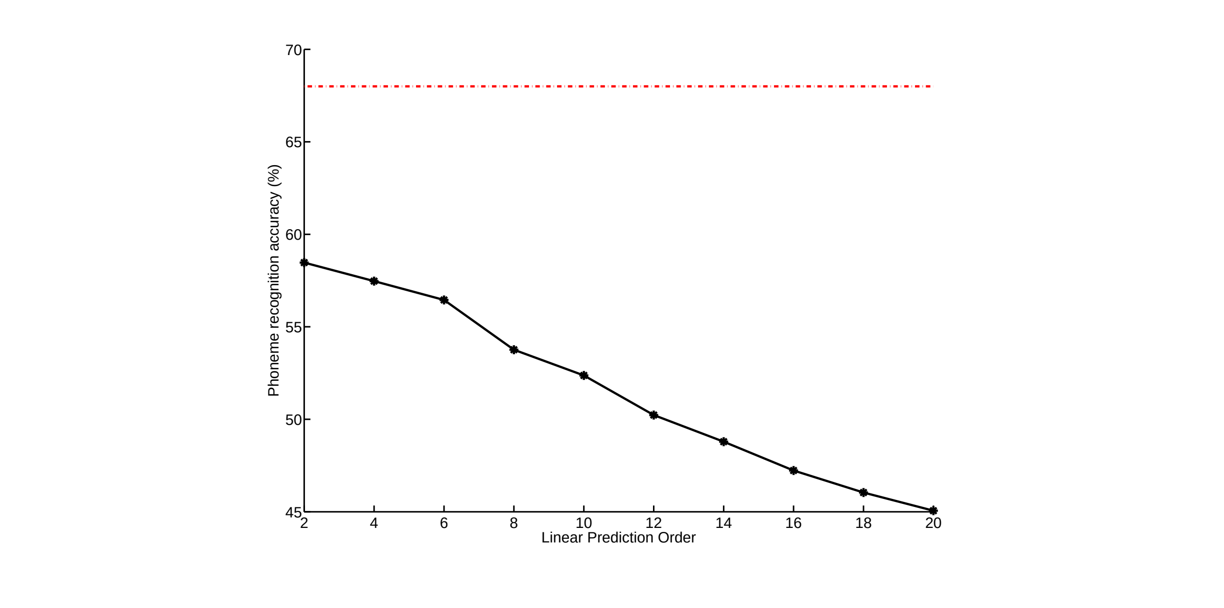

Figure 3.5: Phoneme recognition accuracy for the residual based features various LP orders on TIMIT. The x-axis shows the LP order while the y-axis shows the phoneme accuracy in (%). Phoneme recognition accuracy of reference MFPLP features is shown as a red dotted line.

same, and is derived from the output of the MLP.

3.7.3 Privacy as Phoneme Error Rate

Figure 3.5 plots the recognition accuracies with respect to increasing LP orders using the phoneme recognition system. It can be observed that as the LP order increases the recognition accuracies drop. We note that an increase in LP order by 2 can allow an extra complex conjugate pole pair to be modeled, possibly modeling an extra formant. Since lower order formants generally carry more linguistic information, one could expect the performance to drop when the LP order is increased.

From Figure 3.5, LP residual for a prediction order of 8 has a phoneme recognition accuracy of 53.8%. We remark that the phoneme recognition experiments using simple features, *EZK* and *AH* features, with delta and acceleration coefficients, and a 9 frame context, yielded accuracies of 40.8% and 31.2% respectively. The performance of an 8^{th} order LP residual (53.8%) lies between that of the simple features and the MFPLP features (68.0%).

3.7.4 Enforcing Stricter Privacy Requirements

Table 3.5 lists the phoneme recognition accuracies for obfuscation methods on LP residual and MFPLP features for different block sizes. We note here that randomization can be performed for (a) only test data - second column in the table or (b) both train and test data with different seeds - next two columns in the table. The difference between the two stems from the fact that in the second case, the MLP has been trained with noisy targets. It can be observed that randomized training improves the performance and that as the block size N for randomization increases, the performances of LP residual and MFPLP decrease.

Similarly, we observe from the table that local averaging also provides privacy through a decrease in phoneme recognition accuracies as a function of block size, with randomization providing correspondingly lower phoneme accuracies than averaging. For example, LP residual with 13-frame averaging yields 39.8% while LP residual with 13 randomization yields 29.1% – which is much lower than *EZK* and is also lower than *AH*.

This shows that while linear prediction (with varying prediction orders) provides a degree of control in the allowing linguistic information (privacy), another approach to control the linguistic information can be exploited through temporal randomization or averaging. From the table, it can be seen that obfuscation methods on LP residual yields lower phoneme recognition accuracies than on spectral-shape based features. For this reason, we investigate these methods (randomization, averaging) on LP residual in the next section.

Table 3.5. Phoneme recognition accuracy(%) for MFPLP and LP residual of order 8 for different randomization and averaging block sizes. Linear prediction residual is shown as LPR. Randomization can be performed for (a) only test data - second column or (b) both train and test data with different seeds - next two columns.

Block size (N)	LPR	LPR	MFPLP	LPR
	Clean train Randomized test	Randomized train Randomized test		Averaging
1	53.8	53.8	68.0	53.8
5	42.3	44.1	63.7	50.7
9	33.7	35.1	55.0	45.1
13	28.0	29.1	46.1	39.8
EZK (no randomization): 40.8				
AH (no randomization): 31.2				

3.7.5 Analysis of SND Performance for Obfuscation Methods in Matched Condition

Table 3.6 reports results for obfuscation methods. For the sake of quick reference, we repeat the results of the following SND systems from Table 3.3: $F(LPR8_{51})$, $C(LPR8_{51}, AH_{51})$, $C(LPR8_{51}, EZK_{51}, AH_{51})$, and $F(MFPLP_{31}, EK)$. We now summarize the SND performance under three categories in matched conditions.

(a) Averaging features: Both train and test sets are locally averaged with various block sizes N. It can be observed that for averaging with block sizes N equal to 5, 9, or 13 frames, there is a small drop in performance in comparison with the case where there is no averaging. $C(LPR8_{51}^{A13}, AH_{51})$, denoting the classifier combination of the system trained on a MFPLP representation of 8^{th} order LP residual with 13 frame averaging and the system trained on $F(AH_{51})$, is comparable with the state-of-the-art system, $F(MFPLP_{31}, EK)$.

(b) Randomized {train + test} conditions: In this case, we train randomized features with the correspondingly synchronized groundtruths. The train and test datasets are randomized with different seeds. It can be observed that for 8^{th} order LP residual with a randomization size N equal to 5 or 9 frames, there is no appreciable difference in performance in comparison with no randomization. On the other hand, for a randomization size of 13 frames, there is a small drop in performance.

(c) Clean train condition + randomized test condition: In this case, we use the trained MLP nets on the original unrandomized features with the corresponding unrandomized groundtruths and test them on the randomized test data. On the NIST and ICSI datasets, there is a drop of about 1.5%, which is not substantial in comparison with the performance drop observed in phoneme recognition. Furthermore, in both this case and in the previous case, combination with *AH* features yields state-of-the-art performance.

3.7.6 Analysis of SND Performance for Obfuscation Methods in Mismatched Condition

From the Table 3.6, the performance of the features in mismatched conditions for the obfuscation methods is analyzed under the same categories.

(a) Averaging: It can be observed that for averaging with block sizes N equal to 5, 9, or 13 frames,

Table 3.6. Performance of averaged/randomized test features (in percentage of area under ROC) with a context of 51 frames, in mismatched conditions. The second column lists the overall performance of each system. N, A, and I refer to NIST, AMI, and ICSI datasets. $A \rightarrow B$ refers to the system being trained on a dataset A and being tested on a dataset B. Rx denotes randomization with a block size $N = x$. The table is grouped into blocks of reference, averaged, and randomized (both train and test or test alone) features. For each column and for each block, the best performance is highlighted in bold. The dataset protocol, mentioned in Section 3.3.2, is also mentioned for the respective columns in the table.

Features	All	N	A	I	N→A	N→I	A→N	A→I	I→N	I→A
		NN14	AA25	II36	NA15	NI16	AN24	AI26	IN34	IA35
		Matched conditions			Mismatched conditions					
$F(LPR8_{51})$	84.8	83.0	90.9	89.0	84.5	79.6	83.4	85.3	83.3	87.8
$C(LPR8_{51}, AH_{51})$	**86.3**	86.1	91.8	89.8	88.4	**82.0**	86.4	86.1	85.1	89.8
$C(LPR8_{51}, EZK_{51}, AH_{51})$	86.0	**87.5**	**92.0**	88.9	**88.7**	81.8	**87.8**	85.4	**86.0**	**90.0**
$F(MFPLP_{31}, EK)$	85.0	84.5	91.6	**89.9**	87.4	77.2	86.1	**86.3**	85.3	**90.0**
Averaged features										
$F(LPR8^{A5}_{51})$	84.4	82.4	90.8	89.5	84.8	78.6	82.2	84.9	82.0	87.5
$F(LPR8^{A9}_{51})$	84.2	81.4	90.7	89.2	84.5	78.6	81.5	84.9	81.6	87.1
$F(LPR8^{A13}_{51})$	83.9	81.3	90.4	89.1	83.5	78.6	80.8	83.9	81.3	87.3
$C(LPR8^{A13}_{51}, AH_{51})$	**85.9**	**85.6**	**91.7**	**89.9**	**88.1**	**81.2**	**85.9**	**85.4**	**84.6**	**89.6**
Randomized {train + test} condition										
$F(LPR8^{R5}_{51})$	85.0	83.0	90.7	89.3	83.3	80.6	82.9	85.2	83.0	87.5
$F(LPR8^{R9}_{51})$	83.9	82.3	90.4	88.4	82.9	78.9	81.8	84.3	82.1	87.1
$F(LPR8^{R13}_{51})$	83.3	81.3	90.1	88.0	82.3	77.8	81.5	84.0	81.4	86.2
$C(LPR8^{R13}_{51}, AH_{51})$	**85.7**	**85.6**	**91.5**	**89.4**	**87.8**	**81.0**	**86.0**	**85.6**	**84.5**	**89.3**
Clean train condition + randomized test condition										
$F(LPR8^{C5}_{51})$	84.8	83.1	90.5	89.2	84.4	79.9	82.5	85.0	83.1	87.8
$F(LPR8^{C9}_{51})$	84.0	82.3	90.2	88.6	83.7	78.8	81.9	84.1	82.5	87.4
$F(LPR8^{C13}_{51})$	83.1	81.5	89.9	87.8	82.9	77.7	81.1	83.3	81.8	87.0
$C(LPR8^{R13}_{51}, AH_{51})$	**85.6**	**85.6**	**91.5**	**89.5**	**87.8**	**80.9**	**85.9**	**85.2**	**84.6**	**89.6**

there is a small drop in performance in comparison with the case where there is no averaging, except for the A→N case where there is a drop of 2.6%.

(b) Randomized {train + test} conditions: Unlike the matched case, for a randomization size of 13 frames, there is, in general, a performance drop of about 2%. On the other hand, the drop in performance for the combination with *AH* features is small (less than 1% in all cases).

(c) Clean train conditions + randomized test conditions: Like the matched case there is a performance drop of little more than 2% in many cases. However, combination with *AH* yields comparable performances to unrandomized case (less than 1% in most cases).

Comparing randomization in SND and phoneme recognition, to normalize the advantage that a larger temporal context provides SND in the randomization case, we increased the temporal context of features for phoneme recognition experiments to as much as 51 frames (performed model selection again for this setup). This only decreased the phoneme recognition accuracies. We therefore conclude that randomization affects phoneme recognition much more (around 30%) than it does SND (around 2%).

The second column lists the overall performance of each system. We observe that, for obfuscation methods in general, there is a drop in SND performance for LP residual features. However, this drop in performance is small: for example, 13 frame averaging yields a drop in performance by 0.9%, and a 13 frame randomization yields a drop in performance by 1.7%. For the LP residual systems combined with simple features, this drop is even lesser.

3.8 Final Discussion and Conclusion

Our study investigated three different approaches to privacy-sensitive features for speech/nonspeech detection (SND). These approaches are based on: (a) simple, instantaneous feature extraction methods (b) excitation source information based methods (c) local feature obfuscation methods such as temporal averaging and randomization. To evaluate these features, we used the multiparty conversational meeting data of nearly 450 hours. On this dataset, we evaluated these features and benchmarked them against state-of-the-art spectral shape-based features (MFPLP), on matched and mismatched conditions. To gain further insights, the results were then analyzed for close-talking and far-field microphone scenarios. To quantify the notion of privacy, we conducted phoneme recognition studies on TIMIT. Our investigations are summarized in Table 3.7 and they suggest the following.

Table 3.7. Summary of this chapter in terms of (a) SND performance over all the data measured using AROC (%) and (b) privacy assessment over TIMIT using phoneme recognition accuracy (%).

Features	SND Results AROC (%)	Privacy assessment Phoneme accuracy (%)
Simple features	83.4	40.8
Residual features	86.3	53.8
Randomized residual features	85.6	28.0
State-of-the-art	85.0	68.0

Simple Features

We evaluated the robustness of two sets of simple privacy-sensitive features: (a) energy, zero crossing rate, spectral flatness measure, and kurtosis. (b) Autocorrelation and spectral entropy based features. Explicitly modeling the temporal context is useful for SND in matched and mismatched conditions. For all single features such as energy, zero-crossing, and kurtosis, the increase in performance due to an increase in temporal context is more significant in the far-field case than the near-field case. Furthermore, combinations of single features yield a bigger gain for the far-field case than the close-talking case. Our studies also show that state-of-the-art performance, comparable to MFPLP features, can be achieved by these simple features for the close-talking scenario.

Excitation Source Information

Characterizing the excitation source information using LP residual, we showed that exploiting temporal support of up to 51 frames can yield significant gains in the performance. The residual based feature, while performing at only slightly better levels than simple features on close-talking evaluations, performs significantly better when evaluated on far-field data. We also observed that excitation based features are robust not only with respect to distance, but also with respect to mismatched conditions. Fusion strategies combining LP residual with simple features show that state-of-the-art performance can be obtained in both matched and mismatched conditions, on close-talking and far-field microphone scenarios.

Local Temporal Randomization and Averaging

We investigated the use of local temporal randomization and averaging (up to 130 ms) on the LP residual features. These approaches caused a small drop in SND performance. However, combinations of the randomized or averaged features with simple features yield state-of-the-art SND performance at stricter privacy requirements, defined in terms of phoneme recognition accuracies. These approaches can also be applied to MFPLP features. However, this yields higher phoneme recognition accuracies.

Putting Privacy and SND Performance Together

We quantified privacy in audio through phoneme recognition studies on TIMIT. On the one hand, standard spectral features such as MFPLP yielded, not surprisingly, state-of-the-art phoneme recognition accuracies. On the other hand, simple features yielded much lower phoneme recognition accuracies. LP residual based features yielded phoneme recognition accuracies in between the simple features and the standard spectral features, with the LP order determining the actual performance. Local feature obfuscation methods such as temporal randomization or averaging caused a substantial fall in phoneme recognition performance, with randomization yielding lower phoneme accuracies. SND performance, on the other hand, was relatively unaffected by the temporal obfuscation methods. While it is known that the information in the temporal dynamics of the speech signal can be exploited for phoneme recognition, however, for SND the combination of results showing the importance of temporal context and the relative insensitivity to randomization leads to the conclusion that there is perhaps more information in the statistics of the frames in the temporal support than in the actual temporal dynamics.

Glossary of Notation

Table 3.8: Glossary of notation and their definitions.

Notations	Dim	Definition
		Feature sets
EZK	9	energy, zero-crossing rate, and kurtosis (with delta and acceleration coefficients).
SEZK	12	spectral flatness, energy, zero-crossing rate, and kurtosis (with delta and acceleration coefficients).
AH	9	non-initial maximum of the normalized autocorrelation, the number of autocorrelation peaks and the relative spectral entropy (with delta and acceleration coefficients).
MFPLP	45	MFPLP representation of signal with energy and kurtosis and with delta and acceleration coefficients.
LPR8	39	MFPLP representation of 8^{th} order LP residual with delta and acceleration coefficients.
		MLP based SND systems based on (individual and combinations of features):
$F(E_1)$	3	energy with no context (with delta and acceleration coefficients).
$F(Z_1)$	3	zero-crossing rate with no context (with delta and acceleration coefficients).

Continued on next page

Table 3.8 – continued from previous page

Notations	Dim	Definition
$F(K_1)$	3	kurtosis with no context (with delta and acceleration coefficients).
$F(E_{51})$	153	energy using 51 frame context (with delta and acceleration coefficients).
$F(Z_{51})$	153	zero-crossing rate using 51 frame context (with delta and acceleration coefficients).
$F(K_{51})$	153	kurtosis using 51 frame context (with delta and acceleration coefficients).
$F(EZK_1)$	9	*EZK* features using no context (with delta and acceleration coefficients).
$F(EZK_{51})$	459	*EZK* features using 51 frame context (with delta and acceleration coefficients).
$F(SEZK_{51})$	612	*SEZK* features using 51 frame context (with delta and acceleration coefficients).
$F(AH_{51})$	459	*AH* features using 51 frame context (with delta and acceleration coefficients).
$F(LPR8_{51})$	1989	MFPLP representation of LP residual of prediction order 8 using 51 frame context (with delta and acceleration coefficients).
$F(LPR8_{51}, EZK_{51})$	2448	MFPLP representation of LP residual and *EZK* features using 51 frame context (with delta and acceleration coefficients).
$F(LPR8_{51}, SEZK_{51})$	2601	MFPLP representation of LP residual and *SEZK* features using 51 frame context (with delta and acceleration coefficients).
$F(MFPLP_{31}, \overline{DA})$	403	MFPLP representation of signal with 31 frame context and without delta and acceleration coefficients.
$F(MFPLP_{31})$	1209	MFPLP representation of signal with 31 frame context and with delta and acceleration coefficients.
$F(MFPLP_{31}, EK, \overline{DA})$	465	MFPLP representation of signal with 31 frame context with energy and kurtosis without delta and acceleration coefficients.

Continued on next page

Table 3.8 – continued from previous page

Notations	Dim	Definition
$F(MFPLP_{31}, EK)$	1395	MFPLP representation of signal with 31 frame context with energy and kurtosis and with delta and acceleration coefficients.
$C(LPR8_{51}, EZK_{51})$	1989, 459	combination of $F(LPR8)$ and $F(EZK_{51})$ using equal weights with 51 frame context.
$C(LPR8_{51}, SEZK_{51})$	1989, 612	combination of $F(LPR8)$ and $F(SEZK_{51})$ using equal weights with 51 frame context.
$C(LPR8_{51}, AH_{51})$	1989, 459	combination of $F(LPR8)$ and $F(AH_{51})$ using equal weights with 51 frame context.
$C(LPR8_{51}, EZK_{51}, AH_{51})$	1989, 459, 612	combination of $F(LPR8)$, $F(EZK_{51})$, and $F(AH_{51})$ using equal weights with 51 frame context.
$F(LPR8_{51}^{Ax})$	1989	averaged MFPLP representation of *LPR8* features over a block of x frames using 51 frame context.
$F(LPR8_{51}^{Rx})$	1989	randomized MFPLP representation of *LPR8* features over a block of x frames using 51 frame context (both train and test).
$F(LPR8_{51}^{Cx})$	1989	randomized MFPLP representation of *LPR8* features over a block of x frames using 51 frame context (only test data).

Chapter 4

Speaker Change Detection

Speaker diarization consists of two steps, namely, speaker change detection followed by a clustering of these speaker segments. The next chapter presents a comprehensive investigation of privacy-sensitive features for speaker diarization as a whole. Before that, in this chapter, we will discuss the simpler problem of speaker change detection (SCD). Towards this goal, we shall continue our study of the efficacy of linear prediction (LP) residual for this task.

In summary, this chapter studies privacy-sensitive features based on three different principles: (a) characterizing the excitation source information using the linear prediction residual, (b) characterizing subband spectral information shown to contain speaker information, and (c) characterizing the general shape of the spectrum. We then present experiments on the HUB4 dataset comparing the performance of the privacy-sensitive features to Mel Frequency Cepstral Coefficients (MFCC) features. This work was originally published in Parthasarathi *et al.* (2009b).

The rest of the chapter is organized as follows. The motivation for the selected features is provided in Section 4.1. The experimental setup comprising the dataset, SCD system, baseline features, privacy-sensitive features, and the evaluation measure is described in Section 4.2. Finally, the results, discussion and conclusions are provided in Sections 4.3, 4.4, and 4.5, respectively.

4.1 Motivation and Prior Work

State-of-the-art SCD systems use short-term spectral-shape based features. For instance, the system described in (Ajmera *et al.*, 2004) uses MFCC. These features tend to model the peaks in the spectral envelope, which carry linguistic information. In this regard, speech synthesis studies (Donovan, 1996) have shown that information about the first two formants are important to synthesize intelligible speech. Deriving motivation from SND and phoneme recognition results in the previous chapter, we now investigate the LP residual for SCD.

In addition to LP residual, we shall study two other sources of speaker information, namely (a) higher frequency spectral subbands, and (b) general shape of the spectrum. This section discusses some of the motivations and the prior work for these approaches.

4.1.1 LP Residual

In this section, we shall assume the privacy-sensitive reinterpretation of the LP residual provided in Section 3.1.2. Previous works have shown that the LP residual carries some speaker information (Thevenaz and Hugli, 1995; Dhananjaya and Yegnanarayana, 2007). A key challenge with utilizing LP residual as a feature is to find a suitable representation. One way to represent the LP residual is to estimate its real cepstrum as was done by Thevenaz and Hugli (1995) or to compute the MFCC or PLP coefficients as done in Chapter 3. Other representations of the residual have been explored. For example, Dhananjaya and Yegnanarayana (2007) use the residual without any transformation for a dyadic SCD task, and a group delay representation of the residual was explored by Smits and Yegnanarayana (1995).

While studies using LP residual in the previous chapter provide motivation for a Mel-like or a Bark-like warping of the frequency axis, our aim in this chapter is to explore if the LP residual can be gainfully used for an SCD task. To this end, we shall restrict ourselves to a real cepstral representation of the LP residual. Mel-like frequency warping is investigated for speaker diarization in the next chapter.

Although the usage of a real-cepstral representation of LP residual is similar to the work by Thevenaz and Hugli (1995), the privacy-sensitive setting of this chapter precludes a combination of LP residual with the LP coefficients, as was done by them. On the other hand, if LP residual is used as an independent and stand-alone feature, other sources of speaker information are ignored. Consequently, we endeavor to combine LP residual with speaker information in higher frequency spectral subbands, and in the spectral slope.

4.1.2 Characterizing Subband Information

Previous studies have shown that the spectral subband from 2500 Hz to 3500 Hz carries speaker specific information (Furui, 1986). In this study, we also investigate the two neighboring non-overlapping subbands, namely, 1500 Hz - 2500 Hz and 3500 Hz - 4500 Hz, to assess the importance of the subband 2500 Hz to 3500 Hz. The information in these subbands needs to be suitably represented. We investigate two different representations of the subband information: (a) Computing three MFCC coefficients from the subband. (b) Computing the log-energy from a single filter (centroid) on a subband.

The advantage of the MFCC representation over simple subband filterbank energies is that it decorrelates the filterbank energies and makes these suitable for a Gaussian Mixture Model (GMM) with diagonal covariances. Computing the log-energy of a subband yields a simple representation of the subband information that is suitable for modeling with a Gaussian random variable.

4.1.3 Characterizing Spectral Shape

Speakers differ from each other in the distribution of spectral energies within their speech (Soong and Rosenberg, 1988). Further, it is known that male and female speakers exhibit different spectral energy distribution. In general, the spectra of female speakers show a steeper slope than male speakers. Spectral slope (SS) is thus a way to characterize the shape of the spectrum. In our study, the first cepstral coefficient (c_1) obtained from LP analysis was used as a measure of the spectral slope.

4.2 Experimental Setup

The experimental setup was designed to compare the proposed privacy-sensitive features with the baseline MFCC features. In this section, we describe the dataset, SCD system, baseline, and proposed features used to evaluate the features.

4.2.1 Dataset

The HUB-4 1997 evaluation set was used to test the performance of the proposed features. The HUB-4 database consists of nearly 3 hours of broadcast news data in different acoustic conditions. This data contains a total of 515 speaker changes from a large variety of speakers.

4.2.2 SCD System

In our experiments we compare the baseline features with the proposed features using a state-of-the-art SCD system proposed in (Ajmera *et al.*, 2004). A brief summary of this SCD system is provided below.

Speaker change at a time t in an analysis window is hypothesized by modeling each of the two test subsegments by using a single Gaussian density with the same number of parameters, and by modeling the entire segment with a single GMM. The GMM is modeled with diagonal covariance.

Two neighboring windows are compared using a dissimilarity function based on simplified Bayesian Information Criterion (BIC). This function is computed as the difference between the sum of the log likelihood values obtained from subsegment models and the log likelihood value from the single GMM. A peak value of the distance metric in regions greater than 0 is hypothesized as a speaker change point. Furthermore, it was shown in Ajmera *et al.* (2004) that using the simplified BIC criterion avoids the selection of the threshold used in BIC. It is to be emphasized that this system is kept constant while experimenting with baseline and proposed features.

4.2.3 Baseline Features

The baseline features from Ajmera *et al.* (2004) are 12-dimensional MFCC feature vectors extracted every 10 ms, using a Hamming window of size 30 ms. Similar to the previous work (Ajmera *et al.*, 2004), delta and acceleration features are not used. These baseline features are used with the SCD system described in Section 4.2.2.

4.2.4 Proposed Features

The speech signal is first pre-emphasized (coefficient being 0.97), and then analyzed with a Hamming window of length and shift 30 ms and 10 ms, respectively. The effect of the LP order was investigated by varying the LP order from 4 to 14. A 16th order real cepstrum of the LP residual was estimated. The choice of the cepstral order was based on previous work (Thevenaz and Hugli, 1995). The first cepstral coefficient (c_1) obtained from a 12th order LP analysis was used as a measure of the spectral slope. Three dimensional MFCC feature and log-energy representations of three different subbands, namely, 1500 Hz - 2500 Hz, 2500 Hz - 3500 Hz, and 3500 Hz - 4500 Hz were investigated. The proposed features (up to 21 dimensions) form the input to the SCD system described in Section 4.2.2.

4.2.5 Evaluation Measure

The performance of an SCD system is evaluated based on the two types of errors. A Type-I error is said to occur if the system does not detect a speaker change point within a window. We have used the same size of window as done in (Ajmera *et al.*, 2004), i.e., a window of size 1 second. A Type-II error occurs when a speaker change point is detected but it does not exist in the reference. The Type I and II errors are also evaluated as precision (P) and recall (R) respectively. These are defined as:

$$\mathrm{P} = \frac{\text{number of changes found correctly}}{\text{total number of changes found}} \cdot 100 \quad (\%) \tag{4.1}$$

$$\mathrm{R} = \frac{\text{number of changes found correctly}}{\text{total number of changes}} \cdot 100 \quad (\%). \tag{4.2}$$

In order to compare the performance of different systems, the F-measure is used and is defined as

$$\mathrm{F} = 2 \cdot \frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}} \quad (\%) \tag{4.3}$$

A higher F-measure indicates a better performance.

4.3 Results

The results of all the experiments on the privacy-sensitive features and the baseline MFCC features are reported in Tables 4.1, 4.2, and 4.3 on the HUB-4 1997 evaluation set using precision (P), recall (R), and F-measure (F). In the discussion that follows, LPR-x denotes the 16th order real cepstrum of the residual of LP order x, SS denotes the spectral slope estimated using cepstral coefficient (c_1), MFCC(a - b) denotes the subband MFCC coefficients from a kHz to b kHz, and E(a - b) denotes the subband log-energy value from a kHz to b kHz. The findings of the study are summarized as follows.

4.3.1 Performance of Privacy-Sensitive Features

Table 4.1 compares the performance of the privacy-sensitive features with baseline full-band MFCC features. It can be observed that adding either spectral slope or the subband MFCC to the LP residual cepstrum increases the performance (F-measure). We note that combining spectral slope with LP residual features yields a performance as good as the baseline MFCC features. Combining all the three privacy-sensitive features gives a slight improvement over the baseline MFCC features. It is interesting to note that the SCD system which models the features using Gaussian distributions is suitable for the proposed features as well.

Table 4.1. Complementarity of information in LPR, SS and FB: LPR-x denotes the real cepstrum of LP residual of order x, SS denotes spectral slope, and MFCC(a - b) denotes MFCC values from a kHz to b kHz. The best performance by MFCC baseline is highlighted in bold and italics while the best performances by privacy-sensitive features are highlighted in bold. The dimensions of the 4 feature vectors are 12, 17, 18, 20 and 21 respectively.

Features	P (%)	R (%)	F (%)
MFCC (Baseline)	63.00	64.47	***63.72***
LPR-4	57.98	67.38	61.31
LPR-4 + SS	67.60	60.78	64.00
LPR-4 + MFCC (2.5 - 3.5)	57.14	69.13	62.57
LPR-4 + SS + MFCC (2.5 - 3.5)	66.60	63.50	**65.01**

Table 4.1 shows that baseline MFCC features provide a balance between precision and recall. On the other hand, using residual features by itself yields a higher amount of recall at a lower precision. The addition of subband MFCC to LP residual increases the recall at the same level of precision. Whereas, combining spectral slope with residual features increases the precision. Finally, we observe that combining all the three features results in a more balanced segmentation.

4.3.2 Representing Subband Information

In this section, we investigate (a) the optimal subband, and (b) a representation of subband information for SCD. Table 4.2 shows the performance of three non-overlapping frequency bands represented with MFCC and log-energy values. We note that with either subband MFCC values or with subband log-energies, the subband 2500 Hz to 3500 Hz yields the best performance. This corroborates with earlier studies (Furui, 1986).

Further, we note that the ranking of the three subbands in terms of performance is the same for both subband MFCC and subband log-energy representations. The table also reveals that the subband MFCC representation is a better representation than the subband log-energy representation. In fact, from Tables 4.1 and 4.2 it can be observed that the addition of log-energy value brings down the performance.

Table 4.2. Representing subband information: LPR-x denotes the real cepstrum of the LP residual of order x, SS denotes spectral slope, MFCC(a - b) denotes MFCC values from a kHz to b kHz, and E(a - b) denotes log-energy values from a kHz to b kHz. The first 3 feature vectors have a dimensionality of 21 while the next 3 have a dimensionality of 19.

Features	P (%)	R (%)	F (%)
Representing subband information with MFCC			
LPR-4 + SS + MFCC (1.5 - 2.5)	65.68	60.58	63.02
LPR-4 + SS + MFCC (2.5 - 3.5)	66.60	63.50	**65.01**
LPR-4 + SS + MFCC (3.5 - 4.5)	65.19	60.00	62.48
Representing subband information with log-energy			
LPR-4 + SS + E (1.5 - 2.5)	62.27	58.64	60.40
LPR-4 + SS + E (2.5 - 3.5)	62.23	61.75	**61.99**
LPR-4 + SS + E (3.5 - 4.5)	59.43	61.17	60.29

4.3.3 Effect of LP Order

In this section, we present our investigation on the effect of increasing the LP order. From Table 4.3, it can be observed that increasing the LP order leads to a decrease in the performance up to a prediction order of 10.

We note that an increase in LP order by 2, allows an extra complex conjugate pole pair to be modeled, possibly modeling an extra formant. Since formants carry information about speakers – and LP residual is obtained by filtering out these spectral peaks – we can expect the performance of the LP residual to drop when LP order is increased.

On the other hand, an LP order beyond 10 does not result in a drop in SCD performance. To explain this, we note that the LP residual contains both modeling and excitation errors – where the modeling error corresponds to the all-pole model's error in estimating the vocal tract frequency response. As the LP order increases beyond 10, the contribution of the modeling error in the

Table 4.3. Effect of LP order in LPR: LPR-x denotes the real cepstrum of the LP residual of order x, SS denotes spectral slope, and MFCC(a - b) denotes MFCC values from a kHz to b kHz. All feature vectors have a dimensionality of 21.

Features	P (%)	R (%)	F (%)
Even linear prediction order			
LPR-4 + SS + MFCC (2.5 - 3.5)	66.60	63.50	**65.01**
LPR-6 + SS + MFCC (2.5 - 3.5)	63.62	58.06	60.71
LPR-8 + SS + MFCC (2.5 - 3.5)	63.41	55.53	59.21
LPR-10 + SS + MFCC (2.5 - 3.5)	60.84	50.68	55.30
LPR-12 + SS + MFCC (2.5 - 3.5)	61.47	52.04	56.36
LPR-14 + SS + MFCC (2.5 - 3.5)	59.91	54.56	57.10
Odd linear prediction order			
LPR-5 + SS + MFCC (2.5 - 3.5)	65.39	63.11	**64.23**
LPR-7 + SS + MFCC (2.5 - 3.5)	64.59	56.31	60.17
LPR-9 + SS + MFCC (2.5 - 3.5)	62.01	52.62	56.93

residual signal decreases while the contribution of the excitation error remains constant. In this case, the residual can be likened to modeling the excitation source, which contain speaker information (Plumpe *et al.*, 1999). Experiments performed with LP order approaching 40, showed performance saturating around 60%.

In comparison with an increase in LP order by 2, an increase LP order by 1 does not lead to a big drop in performance. For example increasing LP order from 4 to 5 leads to a drop of only 0.78%. An LP order of 4 can model up to one complex conjugate pole pair, whereas an LP order of 5 can model an extra real pole. Therefore, the performance does not drop much when the LP order is increased from 4 to 5.

4.4 Discussion

We note that the cepstral order of the residual was fixed at 16. However, it would be reasonable to expect the cepstral order to be inversely related to the LP order. For instance, a higher LP order tends to model more formants. Consequently, fewer cepstral coefficients may be sufficient for the LP residual when a high LP order is used.

While this chapter utilizes the real cepstral representation of the LP residual, a number of other representations are possible. For instance, Dhananjaya and Yegnanarayana (2007) used the LP residual directly (without any transformation) for a dyadic SCD task. This was done by modeling the sub-segmental LP residual directly using an auto-associative neural network (AANN) for each speaker. Extensions of this method – particularly, building speaker models – for an unsupervised, multi-speaker segmentation task is not trivial. There have been other representations of the LP residual: for example, the group delay function of the residual signal (Smits and Yegnanarayana, 1995) has been used to determine the instants of significant excitation (corresponding to pitch periods). One of the issues in using group delay functions is that zeros on the unit circle cause it to be ill-behaved. On the other hand, a Mel cepstrum representation of the LP residual cepstrum, as studied in Chapter 3 with much promise, can be explored.

4.5 Conclusions

In this chapter, we investigated the linear prediction (LP) residual in conjunction with subband MFCC (a bank of 4 filters) and the spectral slope, for an SCD task. Using F-measure as an evaluation measure on the HUB-4 1997 evaluation set, experiments showed that the performance of the proposed privacy-sensitive features is comparable or better than that of the state-of-the-art full-band MFCC features. In addition, it was shown that SCD performance was sensitive to LP order. Overall, our study shows that LP residual could be a promising feature for speaker diarization in a privacy-sensitive setting.

Chapter 5

Speaker Diarization

Supported by the phoneme recognition results, we have, so far, interpreted the LP residual as privacy-sensitive and have investigated it for speech/nonspeech detection (SND) and speaker change detection (SCD). In Chapter 3, we concluded that LP residual in conjunction with simple features yields a state-of-the-art SND performance. On the other hand, in Chapter 4 for a speaker change detection task, we introduced two other sets of features, namely higher frequency spectral subbands, and the spectral slope. We studied the combination of LP residual with these features, showing that a performance comparable to that of Mel Frequency Cepstral Coefficients (MFCC) can be achieved. Speaker diarization, however, is a more complex task and it involves an additional step, namely speaker clustering.

This chapter endeavors to explore privacy-sensitive methods for speaker diarization. The source-filter model assumed by linear prediction provides a natural framework for exploring privacy concerns. As an extension, data-driven methods assuming the source-filter model (not including linear prediction) is explored. The notion of privacy preservation while capturing speaker information is discussed in more formal terms using Mutual Information (MI).

In this chapter, we shall discuss speaker diarization in single and multiple distant microphone scenarios, systematically investigating the Linear Prediction (LP) residual. Issues such as prediction order and choice of representation of LP residual are studied. Additionally, we explore the combination of LP residual with subband information from 2.5 kHz to 3.5 kHz and spectral slope. Next, we propose a supervised framework using deep neural architecture for deriving privacy-sensitive audio features. We benchmark these approaches against the traditional MFCC features for speaker diarization in both the microphone scenarios. Experiments on the RT07 evaluation dataset show that the proposed approaches yield diarization performance close to the MFCC features on the single distant microphone dataset. To objectively evaluate the notion of privacy in terms of linguistic information, we perform human and automatic speech recognition tests, showing that the proposed approaches to privacy-sensitive audio features yield much lower recognition accuracies compared to MFCC features. Part of this work is published in Parthasarathi *et al.* (2011a).

The rest of the chapter is organized as follows. Section 5.1 reviews the literature on deep neural networks. The overall methodology of this chapter is summarized in Section 5.2. A description and an analysis of the proposed features is given in Section 5.3, while Section 5.4 discusses the diarization setup. Parameters selection experiments associated with the proposed features are described in Section 5.5. Subsequent validations on an additional dataset are presented in 5.6. We revisit privacy in Section 5.7, and the conclusions are drawn in Section 5.8.

5.1 Related Work

Related works on LP residual motivating privacy preservation as well as retention of speaker information have been discussed in Chapters 3 and 4.

In this section, we summarize relevant work in deep neural networks. In particular, we review the relevant literature on deep neural networks as a means to represent phoneme information. In subsequent sections, we describe privacy-sensitive features derived from a deep neural network architecture.

Multilayer feedforward neural networks with a 3-layer architecture, also called multilayer perceptrons (MLP), have been used for feature extraction in the automatic speech recognition (ASR) community for several years (Bourlard and Morgan, 1994; Hermansky *et al.*, 2000). Recently, deep neural networks, i.e., typically the number of layers being more than three (alternatively, number of hidden layers being more than one), have been receiving attention from both the machine learning and the speech recognition community (Hinton and Salakhutdinov, 2006; Grezl *et al.*, 2007) due to their ability to represent knowledge compactly and in a principled fashion. The motivation for this has been attributed to results from complexity theory of circuits (Larochelle *et al.*, 2009).

Of particular interest to our work are deep neural networks with a bottleneck architecture to represent phoneme information. In the field of automatic speech recognition (ASR), deep neural networks with a bottleneck architecture have recently begun to be investigated in the quest towards obtaining better phoneme representation before further processing by a HMM/GMM system (Grezl *et al.*, 2007). For example, in (Grezl *et al.*, 2007) the output (before the sigmoid nonlinearity) taken from the bottleneck layer of a trained five-layer MLP, was used in a conventional HMM/GMM system to yield promising results.

In addition, deep neural networks with a bottleneck architecture have the advantage of providing the ability to train successive layer of weights to optimize different cost functions. In this chapter, we investigate one such strategy, where the first two layers of the network weights are trained for phoneme classification, while the next two layers of the network weights are trained for a reconstruction of the input. Alternatively, one could think of this as an estimate of the power spectrum of a frame of speech derived from the bottleneck layer of a phoneme MLP.

However, a key issue in exploiting deep neural networks is the inherent difficulty in training the weights. A gradient-based optimization starting from random initialization has been reported to get trapped in local optima leading to poor solutions (Larochelle *et al.*, 2009). This was also observed by us in our studies in training neural networks with more than three layers for phoneme recognition on TIMIT, to the extent that deeper networks perform worse than MLPs with one hidden layer.

Two common strategies to address this difficulty are greedy layer-by-layer training (Bengio *et al.*, 2006) or an autoencoder training (Hinton and Salakhutdinov, 2006). In (Frankel *et al.*, 2008), features derived from the bottleneck layer of a 5-layer deep neural network trained with a greedy layer-by-layer method was shown to yield promising performance for an ASR task on over 100 hours of meeting audio data.

The constraints of privacy in features imply the necessity to capture the complement of phoneme information captured by the bottleneck layer of a 5-layer MLP. In this context, our work exploits features derived from the bottleneck layer of a deep neural network as information that needs to be filtered from the spectrum. In Section 5.3.2, we describe the proposed method in detail.

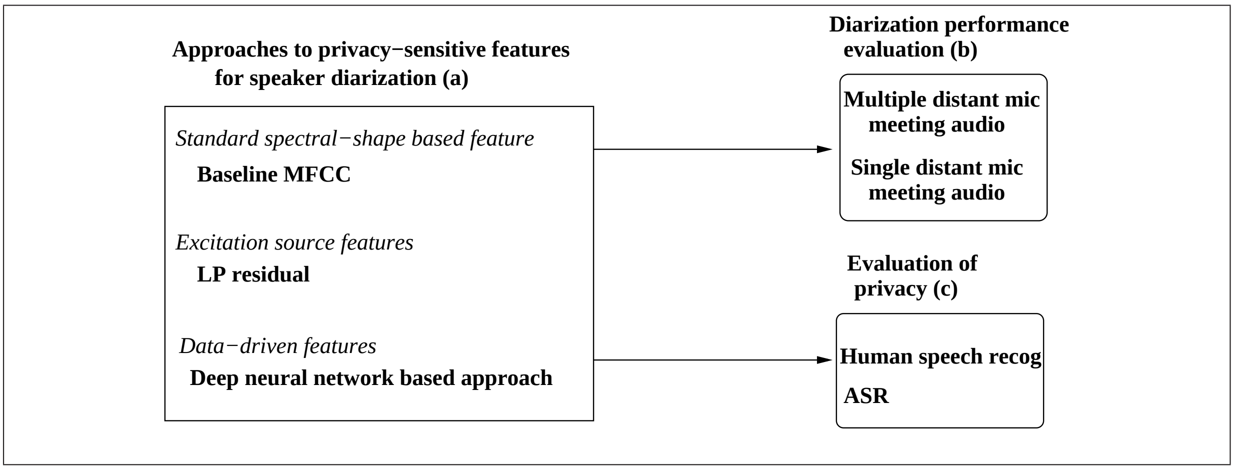
Approaches to privacy−sensitive features for speaker diarization (a)
Standard spectral−shape based feature
Baseline MFCC
Excitation source features
LP residual
Data−driven features
Deep neural network based approach
Diarization performance evaluation (b)
Multiple distant mic meeting audio
Single distant mic meeting audio
Evaluation of privacy (c)
Human speech recog
ASR

Figure 5.1: Block diagram of our approach. A detailed discussion of the figure is provided in Section 5.2.

5.2 Our Methodology

In this section, we summarize our overall methodology, also illustrated using a block diagram in Figure 5.1. These blocks are described below.

(a): We begin with a detailed description of the features extracted from LP residual and deep neural networks. Sections 5.3.1 and 5.3.2 describe these features in detail. To gain insight into the features, this is followed by a more formal analysis of the proposed features in terms of mutual information.

(b): Evaluating privacy-sensitive features entails a comparison of diarization performance as well as an evaluation of linguistic privacy. Details of the diarization system, features, datasets, and the baseline performance figures are presented in Section 5.4. Parameter selection experiments associated with the proposed features for diarization is done on the development data (RTeval06) on single and multiple distant microphone data (Section 5.4). Results on evaluation data (RTeval07) is presented in Section 5.6.

(c): This chapter quantifies linguistic privacy using human listening tests and automatic phoneme recognition studies. Section 5.7 provides further details on the methodology followed and the results obtained using these tests.

5.3 Privacy-Sensitive Features

In this section, we summarize the details in deriving the proposed features. This is followed by an analysis based on mutual information framework.

5.3.1 LP residual based Features

We now look at extracting features from LP residual, subband information, and spectral slope. We then discuss obfuscation strategies.

(a) LP Residual: The LP residual is extracted every 10 ms, using a Hamming window of size 30 ms. The representations of the residual studied are: a real-cepstrum representation (Thevenaz and Hugli, 1995) with a fixed number of 19 coefficients and a MFCC representation with 19 coefficients. The MFCC representation is computed using HTK (Young *et al.*, 2000). These representations have been fixed at 19 dimensions so as to have the same dimensions as the baseline MFCC features. Feature selection experiments investigating the choice of representation is presented in detail in Section 5.5. We then study LP residual by varying the prediction order from 2 to 20. The choice of the LP order presents a tradeoff between privacy and SND performance.

(b) Subband Information: Recall that previous studies have shown that the spectral subband from 2500 Hz to 3500 Hz, carries speaker specific information (Furui, 1986). In Chapter 4, we exploited the relative importance of the subband 2500 Hz to 3500 Hz over the two neighboring subbands (1500 Hz - 2500 Hz and 3500 Hz - 4500 Hz) for a speaker change detection (SCD) task. We also showed that computing three MFCC coefficients from this subband was better than computing the logarithmic energy from the subband. As remarked in the last chapter, a further advantage of the MFCC representation is that it decorrelates the filterbank energies and makes it suitable for a Gaussian Mixture Model (GMM) with diagonal covariance matrices.

(c) Spectral Shape: We recall from the last chapter that speakers differ from each other in the distribution of spectral energies within their speech signals (Soong and Rosenberg, 1988). Spectral slope (SS) is a way to characterize the shape of the spectrum, and in Chapter 4 we showed that the first cepstral coefficient (c_1) obtained from LP analysis can enhance SCD when combined with the LP residual features.

(d) Obfuscation/Local Temporal Randomization: We recall this procedure from Chapter 3. Fea-

ture vectors within a block of size ($N = 1, 5, 9, 13$) are shuffled. A uniform pseudo-random number generator was used to shuffle the frames in the block. It can be noted that a randomization of N frames could result in two successive frames being separated by $2 \cdot (N - 1)$ frames. We chose block sizes up to 13 frames since results in (Pinto *et al.*, 2011) indicate that phonetic information in the speech signal up to 230 ms can be exploited for phoneme recognition.

5.3.2 Deep Neural Network based Features

The aim of the proposed approach is to model the peaks in the spectral envelope that tend to carry linguistic information. For this, the spectral envelope is reconstructed from a phoneme representation. The reconstructed envelope is then filtered to obtain a residual (similar to LP residual), which is represented using MFCC. Details of the two steps – reconstructing the envelope and filtering – and an example, are provided below.

Reconstructing spectral envelope

Reconstruction of the spectral envelope is accomplished in two further steps. First, we train a 5-layer phoneme MLP, with a bottleneck architecture, that performs phoneme classification. From Frankel *et al.* (2008); Grezl *et al.* (2007), output at the bottleneck layer (i.e., bottleneck features) can be considered as a good phoneme representation. As a second step, the output from the bottleneck layer of the phoneme MLP is used to train a reconstruction MLP, which reconstructs the spectral envelope. An illustration of this is provided in Figure 5.2. We now discuss the architecture and training procedure of the two MLPs in detail.

Phoneme MLP: Two phoneme classification MLPs are trained without explicit temporal context. These MLPs take as input either MFCC or logarithm of DFT square magnitude vectors (obtained from 512 point FFT), both of which are mean and variance normalized. When there is no ambiguity, we refer to both of them as *phoneme MLP*. Let the layers of the phoneme MLP and their notations be – input (I), first expansion (H1), bottleneck (B), second expansion (H2), and output (O1). The number of nodes in H1 and H2 was kept same, since experiments in Grezl and Fousek (2008) show that varying the ratio of H1 to H2 did not yield an appreciable difference in ASR performance. The bottleneck layer is a dimensionality reduction layer Grezl *et al.* (2007), and we varied the number of units from 20 to 40 Grezl and Fousek (2008).

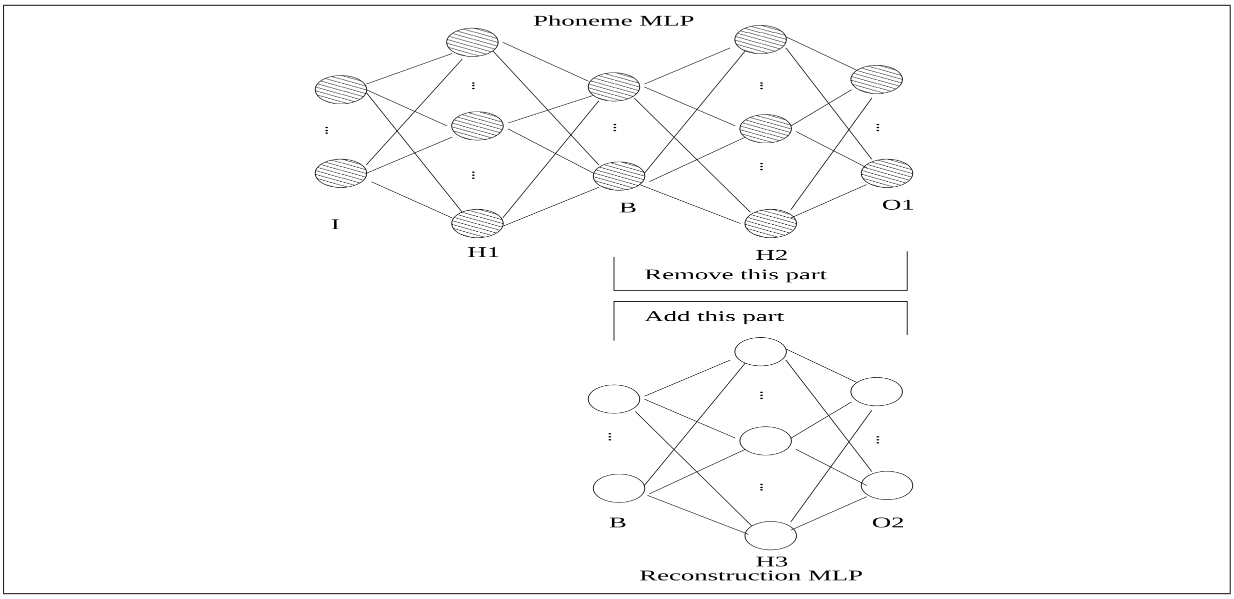

Figure 5.2: 5-layer deep neural network with bottleneck architecture. (a) 5-layer phoneme MLP is trained with phoneme targets using cross entropy criterion (b) Keeping weights for the first 2 layers fixed, and removing last 2 layers, a reconstruction MLP is trained for the last two layers with squared error criterion.

The output layer of the phoneme MLP represents the phoneme class and we use 39 units with softmax nonlinearity. This MLP was trained by growing MLPs layer-by-layer on the TIMIT database Bengio *et al.* (2006). Cascaded MLPs with 3, 4, and 5 layers are trained using standard back propagation algorithm by minimizing the cross entropy error criterion (Frankel *et al.* (2008); Grezl *et al.* (2007)). Excluding 'sa' dialect sentences, the TIMIT training data consists of 3000 utterances from 375 speakers and the cross-validation data consists of 696 utterances from 87 speakers. The hand-labeled dataset using 61 labels is mapped to the standard set of 39 phonemes Pinto *et al.* (2011).

Reconstruction MLP: To reconstruct the spectral envelope, we train a 3-layer regression MLP that takes the bottleneck features as input and reconstructs the power spectrum by minimizing the squared error. The parameters of the reconstruction MLP are: the input from the bottleneck layer (B), the expansion layer (H3), and the output layer (O2).

The input to the reconstruction MLP is the linear output from the bottleneck layer of the phoneme MLP. The number of nodes in the expansion layer (H3) is varied independent of H1 and H2. The output of the reconstruction MLP is the estimated power spectrum, i.e., logarithm of 257 point DFT square magnitude vectors. Another choice of output, namely, a 19 dimensional MFCC was

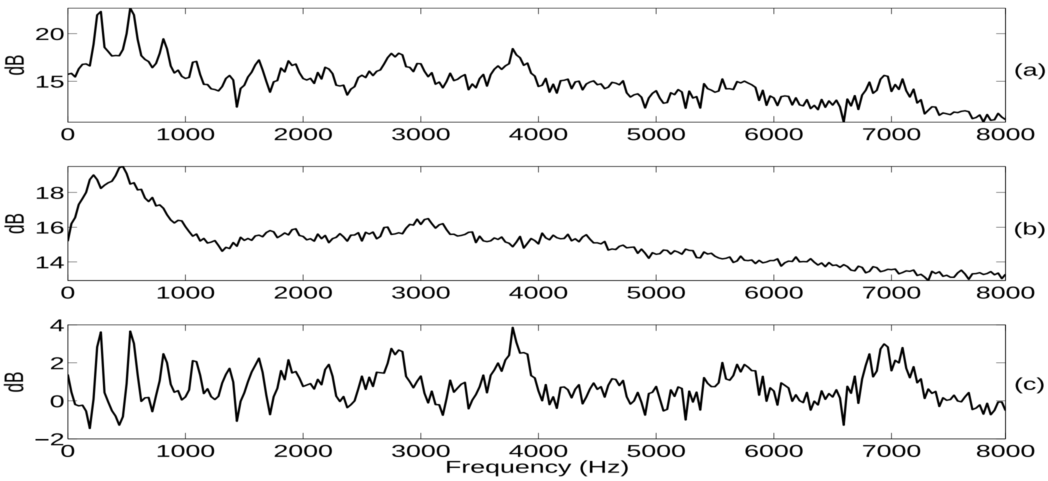

Figure 5.3: Example steps in neural network filtering for an input frame that is /iy/: (a) Input to phoneme MLP (logarithm of DFT square magnitude vector) (b) Output from reconstruction MLP (logarithm of DFT square magnitude vector) (c) Filtered spectrum.

explored. We refer to both MLPs as *reconstruction MLP*. These MLP are trained on TIMIT train set, described above, using standard back propagation algorithm by minimizing the squared error criterion.

Filtering to remove spectral envelope

For an input, MFCC or logarithm of DFT square magnitude vectors, the corresponding phoneme MLP is used to obtain the linear output from the bottleneck layer. Parameter selection experiments are performed with both reconstruction MLPs. The estimated envelope, obtained from the output of the reconstruction MLP, is either logarithm of 257 point DFT square magnitude vectors or 19 dimensional MFCC.

Filtering is then performed to remove the estimated envelope from the original spectrum of the speech signal. For the case where the output units are logarithm of DFT square magnitude vectors, filtering is performed by subtracting it from the input (logarithm of DFT square magnitude vectors). The filtered squared magnitude vector is then converted to an MFCC representation of 19 dimensions. In the case of the output units being MFCC, filtering is performed by subtracting it from the input MFCC.

An example

Figure 5.3 illustrates the example steps in neural network filtering for an input frame that is /iy/ phoneme. Figure 5.3(a) plots the input to the phoneme MLP (logarithm of DFT square magnitude vector). Observe that the broad spectral shape and the spectral details are manifest. First formant can be seen around 320 Hz, while the second formant can be observed around 2500 Hz. Figure 5.3(b) shows the output from reconstruction MLP (logarithm of DFT square magnitude vector). It can be observed that the reconstructed spectrum consists mainly of the spectral shape than the spectral details. Figure 5.3(c) shows the filtered spectrum. We observe that the spectral shape (mainly the first formant) is filtered.

5.3.3 Mutual Information based Analysis

In this section, we present an analysis of the privacy-sensitive features using mutual information. Privacy in audio could be interpreted as a function that maximizes the mutual information (MI) with speakers while minimizing the MI with linguistic information. This framework is discussed next followed by an analysis of the features on TIMIT test data (consisting of 1344 utterances from 168 speakers).

MI Framework

Given X, a multivariate continuous random variable denoting the log squared magnitude, and S, Q discrete random variables, denoting speaker and phoneme labels respectively, the goal is to find a transformation g that maximizes the function $I(g(X); S) - I(g(X); Q)$.

$$g^* = \arg\max_g I(g(X); S) - I(g(X); Q) \tag{5.1}$$

This equation is in general difficult to solve without additional constraints or assumptions. We note here that one can not increase information by processing a signal, one can only remove information from it. On the other hand, such a transformation (g) one could decrease information about the speakers. In this context, we seek that transformation that maximizes Equation 5.1.

Now, assuming that Q and S are independent – it might be that speakers can have biases towards choices of words and therefore towards phoneme – the maximum of (5.1) is reached for:

$$g^*(X) = \tilde{S} \tag{5.2}$$

where $\tilde{S}$ is a transformation of X that has maximum mutual information with S. We now make the second, stronger assumption: i.e., the source-filter model of speech production leads to a separation of the speaker and the phoneme information in the signal space. In general this assumption is not true, as the source-filter model only assumes a separation of the excitation source and the vocal tract response. But one could argue that the separation of broad spectral shape and one consisting of the remainder (e.g. high and low order cepstral coefficients) have more information about the phonemes and speakers, respectively. Mathematically, making this assumption is equivalent to,

$$g^*(X) = \tilde{S} = X - \tilde{X} \tag{5.3}$$

where $\tilde{X}$ is a transformation of X that has maximum mutual information with Q.

LP Residual: In the case of LP, an independent source-filter model assumption is part of the modeling. The all-pole model can be reinterpreted as an estimate of the phoneme information ($\tilde{X}$) and it is obtained in an unsupervised fashion as the smoothed spectral envelope. The LP residual naturally becomes $g^*(X)$ in 5.3.

Deep Neural Network Filter: An alternative is to train a data-driven filter that yields $\tilde{X}$, given X as input. We shall show this. Let us consider a 5-layer MLP for phoneme classification, with a bottleneck architecture. Let X denote the input, and let Z denote the random variable at the output of the MLP. Then,

$$Z = \psi(X; \theta_1, \theta_2, \mathcal{D}) \tag{5.4}$$

where θ_1, θ_2 is the set of all parameters of the MLP (i.e., the weights and the biases) before and after the bottleneck layer respectively, and $\mathcal{D}$ is the training data. Let q_k denote the k^{th} phoneme and $\tilde{P}$

denote the estimated probabilities. The cross-entropy training criterion can be written as:

$$
\begin{aligned}
\mathcal{J}(\theta_1,\theta_2) &= -E_X\Big(\sum_k P(q_k|x)\log\tilde{P}(q_k|x)\Big)\\
&= -\int_X p(x)\sum_k P(q_k|x)\log\tilde{P}(q_k|x)dx\\
&= -\int_X\sum_k P(q_k,x)\log\frac{\tilde{P}(q_k|x)\tilde{P}(x)\tilde{P}(q_k)}{\tilde{P}(x)\tilde{P}(q_k)}dx\\
&= -\int_X\sum_k P(q_k,x)\Big(\log\frac{\tilde{P}(q_k,x)}{\tilde{P}(x)\tilde{P}(q_k)}+\log\tilde{P}(q_k)\Big)dx\\
&= -\int_X\sum_k P(q_k,x)\log\frac{\tilde{P}(q_k,x)}{\tilde{P}(x)\tilde{P}(q_k)}-\int_X\sum_k P(q_k,x)\log\tilde{P}(q_k)dx\\
&= -\int_X\sum_k P(q_k,x)\log\frac{\tilde{P}(q_k,x)}{\tilde{P}(x)\tilde{P}(q_k)}-\sum_k\log\tilde{P}(q_k)\int_X P(q_k,x)dx\\
&= \tilde{I}(Q;X)-\sum_k P(q_k)\log\tilde{P}(q_k) \qquad (5.5)
\end{aligned}
$$

Since the second term on the right hand side of Equation 5.5 does not depend on X, it can be concluded that minimum cross-entropy training is equivalent to maximum mutual information training (Bridle, 1990). Let B denote the random variable obtained at output from the bottleneck layer before the nonlinearity. Then,

$$B=\phi(X;\theta_1,\mathcal{D}) \qquad (5.6)$$

where θ_1 is the set of parameters of the MLP up to the bottleneck layer. Furthermore, from data-processing inequality (Cover and Thomas, 1991),

$$I(X;Q)\geq I(B;Q)\geq I(Z;Q) \qquad (5.7)$$

However, given the constraints of the parameters (θ_1,θ_2), $I(Z;Q)$ is maximized. Similarly, $I(B;Q)$ is maximized for θ_1. This together with the fact that the dimension of the output at the bottleneck (B) is much smaller than that of the dimension of input (X), means that bottleneck (B) serves as a compression of input (X) retaining information that has maximum mutual information with the phonemes (Q).

Therefore, it is reasonable to assume that as the dimension of B is made much smaller than

X, other information such as speakers (S) is lost at bottleneck (B). We now consider the second MLP, namely, the reconstruction MLP: i.e., this MLP is trained taking bottleneck output (B) as input and X as the training target, with minimizing the least-squares error cost function. The random variable at the output of this MLP ($\tilde{X}$) is a reconstruction of X and has therefore the same dimension as X. It is, however, reconstructed using B, which has maximum mutual information with Q (and has low MI with S, because of dimensionality reduction at B). Therefore, $\tilde{X}$ can be considered to be an estimate of Q. Inserting $\tilde{X}$ so obtained in (5.3), we obtain $\tilde{S}$.

MI Analysis

In practice, we can introduce a variable (λ) in (5.1) to make it $I(g(X); S) - \lambda \cdot I(g(X); Q)$ and tune this variable for optimal values. Alternatively, we could plot $I(X; Q)$ versus $I(X; S)$ and make more qualitative assessments on the tradeoff between privacy and speaker information, in using these features. In this chapter, we take the latter approach. Figure 5.4 shows such a plot. That is, $I(X; Q)$ versus $I(X; S)$, on the TIMIT test set. A higher $I(X; Q)$ could be interpreted as a feature with lower privacy. Similarly, a feature yielding higher $I(X; S)$ could be interpreted as a better feature for diarization. An ideal privacy-sensitive feature would be in the top-left of this plot.

For estimating the MI with phoneme and speaker labels, we use the following form of MI: $I(X; A) = H(X) - H(X|A)$, where A denotes either Q or S. To estimate entropies $H(X)$ and $H(X|A)$, we use k-means clustering algorithm to discretize the feature space. The features are then binned and the normalized bin-counts are then used to estimate $I(X; A)$. Model selection on the TIMIT training data is used to identify the number of clusters. Bias correction is performed using the Miller's formula on the estimated mutual information (Miller, 1954).

Figure 5.4 plots baseline MFCC, residual, and deep neural network features represented as 19 dimensional MFCC. Baseline MFCC has high $I(X; S)$, showing that it is a good feature for speaker recognition; on the other hand, it is not privacy-sensitive since it has high $I(X; Q)$. For the residual, it can be observed that as the LP order increases, $I(X; Q)$ and $I(X; S)$ decrease. Clearly, a high LP order yields a privacy-sensitive feature, but it also yields low speaker information. LP order thus offers a tradeoff between privacy and speaker information. A prediction order of 8 seems appropriate since it yields less MI with phonemes than does the baseline MFCC. Furthermore, it would lead to the loss of the first 2 to 3 formants that are important for synthesizing intelligible

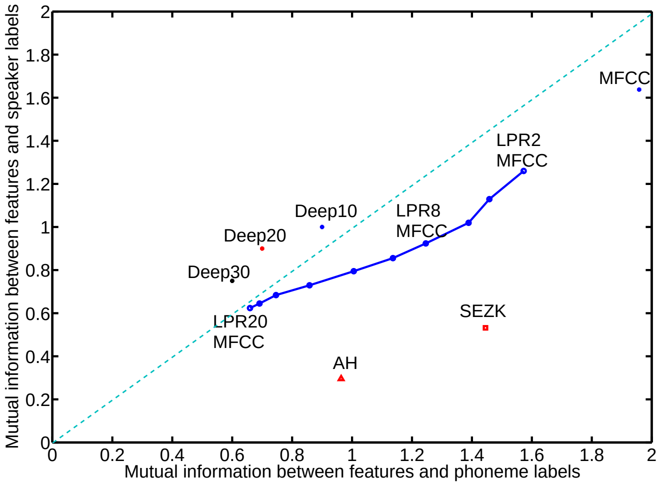

Figure 5.4: Plot showing mutual information between the features and phonemes versus mutual information between the features and speakers. LPRx denotes residual features with LP order x. *SEZK* and *AH* denote the features from (Parthasarathi *et al.*, 2010) and (Wyatt *et al.*, 2007a) respectively. Deepxy refer to deep neural network based features with bottleneck sizes corresponding to xy.

speech Donovan (1996). The figure also plots a Y=X line so that we can quickly judge the slopes. We can observe that only for LPRx with $x > 6$ does the LP residual begin to lose phoneme information faster than speaker information.

For the deep neural network features, the input and reconstruction layers are squared magnitude vectors, with 3 bottleneck sizes ($B = 10, 20, 30$). The expansion layers were fixed at 1000. Similar to the LP order, the number of bottleneck units presents a tradeoff between privacy and speaker information. Having more units enables the capture of the spectral envelope better; however, at the cost of speaker information. In comparison with an eighth order residual, it can be seen that the deep neural network features (with 20 bottleneck units) yield much lower MI with phoneme labels, while yielding similar MI with speaker labels.

Features from Parthasarathi *et al.* (2010) and Wyatt *et al.* (2007a) are marked *SEZK* and *AH*, respectively. *SEZK* is used to denote the feature formed by concatenating spectral flatness, energy, zero crossing rate, and kurtosis; while *AH* denotes a concatenation of non-initial maximum of the normalized autocorrelation, number of autocorrelation peaks, and relative spectral entropy. These features, *SEZK* and *AH*, are privacy-sensitive but have low MI.

5.4 Diarization Setup

This section discusses the baseline system, features, datasets and the performance measure used to evaluate the features.

5.4.1 Baseline Diarization System

The baseline system is an ergodic HMM as described in (Ajmera and Wooters, 2003). Each HMM state represents a cluster (speaker). The state emission probabilities are modeled by Gaussian Mixture Models (GMM) with a minimum duration constraint of 3 seconds. The algorithm follows an agglomerative framework, i.e, it starts with a large number of clusters (hypothesized speakers) and then iteratively merges similar clusters until it reaches the best model. After each merge, data are re-aligned using a Viterbi algorithm to refine speaker boundaries. The initial HMM model is built using uniform linear segmentation and each cluster is modeled with a 5 component GMM. The algorithm then proceeds with bottom-up agglomerative clustering of the initial cluster models (Chen. and Gopalakrishnan, 1998). At each step, all possible cluster merges are compared using a modified version of the BIC criterion (Ajmera and Wooters, 2003).

The diarization system uses 19 dimensional MFCC features and the time delay of arrival (TDOA) features from the beamformed signal. The MFCC vectors are extracted every 10 ms, with a hamming window of size 30 ms, using HTK (Young *et al.*, 2000). Delta and acceleration features are not used.

5.4.2 Privacy-Sensitive Features

The proposed privacy-sensitive features are compared against the baseline 19 dimensional MFCC using the system discussed in Section 5.4.1. To summarize Section 5.3, LP residual is represented using MFCC or real-cepstrum, both 19 dimensional. The 2.5 kHz to 3.5 kHz subband (SB) is represented using 3 dimensional MFCC and is concatenated with the spectral slope (SS), represented using the first cepstral coefficient (c_1) obtained from LP analysis. The two feature streams, one consisting of LP residual and another of SB and SS features, are modeled with different GMMs and they are combined by linearly weighting the individual log-likelihoods Ajmera and Wooters (2003).

For obfuscation, features are shuffled with a uniform random number generator for block sizes ($N = 5, 9, 13$). The deep neural network features are represented using 19 dimensional MFCC.

5.4.3 Datasets

Experiments were performed on NIST RT06 and RT07 evaluation data for Meeting Recognition Diarization task (NIST RT06, 2006; NIST RT07, 2007). RT06 evaluation data is used as the development dataset and it contains nine meeting recordings of approximately 30 minutes each. The best set of parameters is then used for benchmarking the proposed features against MFCC features on the RT07 dataset using the baseline diarization system. The evaluation dataset (RT07) contains eight meetings of nearly 43 minutes each. MDM data is obtained by denoising the individual channels using Wiener filter and then beamforming using the BeamformIt toolkit (Anguera, 2006). SDM experiments were performed on randomly selected individual MDM channels.

Speech/nonspeech segmentation is obtained using a forced alignment of the reference transcripts on close talking microphone data using the AMI RT06 first pass ASR models (Hain *et al.*, 2006). Since our interest in this chapter is in evaluating the privacy-sensitive features for speaker segmentation and clustering, the same speech/nonspeech segmentation is used across all experiments.

5.4.4 Baseline Performance

The results are reported in terms of Diarization Error Rates (DER). DER is the sum of speech/nonspeech errors and speaker errors. Speech/nonspeech errors is the sum of missed speech and false alarm speech. For all experiments reported in this chapter, we include the overlapped speech in the evaluation.

Table 5.1 lists the performance of the baseline diarization system on RT06 MDM and SDM evaluation data. The first 3 columns list the performance of the speech/nonspeech detection system in terms of missed speech, false alarm, and over all speech/nonspeech detection error. The overall speech/nonspeech error rate over all the files on the RT06 evaluation dataset is 6.6%. The next two columns list the performance of the baseline MFCC features in terms of the speaker error for both the MDM and the SDM scenarios. As expected, MFCC features perform better on the development MDM data. On RT06 we observe a performance gain of 3.7% on the MDM data over the SDM data.

Table 5.1. RT06 evaluation data: The first 3 columns list the performance of the speech/nonspeech detection while the next 2 columns report performance of baseline MFCC features for MDM and SDM.

Evaluation	Miss	FA	Sp/nsp	Spkr err (%)	Spkr err (%)
				MDM	SDM
RT06	6.5	0.1	6.6	**17.1**	**20.8**

5.5 Parameter Selection on RTeval06

Recall that we use RTeval06 as the development dataset. In Section 5.3.3, we presented an analysis of the features using MI on the TIMIT test set. In this section we perform parameter selection experiments for the proposed features using the baseline diarization system on RTeval06.

5.5.1 LP Residual based Features

We address three issues in this section: (a) the choice of representation (b) prediction order (c) combination with slope and subband energies.

Representations of LP Residual

We study the 2 different representations of LP residual using the baseline diarization system described in Section 5.4.1. Figure 5.5 shows the comparison between the 2 representations on the RT06 MDM evaluation data. It can be observed that MFCC representation yields a better performance for all prediction orders. It is interesting to observe that the gap between the two representations decrease as the prediction order increases. It could be due to MFCC being better able to capture spectral peaks than real cepstrum. From here on, we use MFCC representation of the residual.

Prediction Order

The effect of LP order on MFCC representation of residual on both MDM and SDM data is presented in Figure 5.6. Both curves exhibit similar behaviors, which can be analyzed separately in 3 relatively distinct regions: smaller drop in performance for increases in prediction orders from 2 to 6, followed by a more dramatic drop in performance for prediction orders between 8 to 12, and then again a smaller drop afterward.

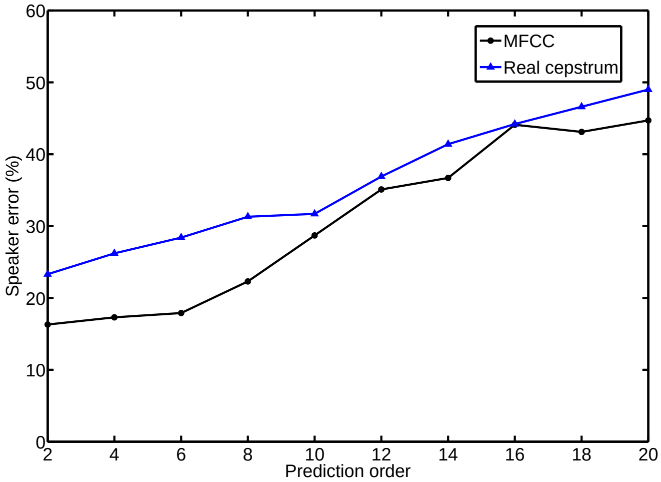

Figure 5.5: Comparison between MFCC and real-cepstrum representations of the LP residual on the development dataset (RT06).

Let us consider prediction orders between 2 to 6. An increase from 2 to 6 results in a drop of 1.6% in the MDM case. This could be due to the loss of the first formant, which carries more linguistic information (Donovan, 1996). Speaker error, therefore, seems to be relatively less affected.

For LP orders between 8 to 12, an increase in the LP order results in a bigger drop in performance. For instance, an increase in LP order from 8 to 10 results in a drop of nearly 6% in MDM and 5% in SDM. We note that the vocal tract system is typically characterized by up to five resonances in the 0 to 4 kHz range. An LP order in the range 8 to 12 can model around 3 formants. Since higher order formants carry more speaker information, we note that increasing prediction order beyond 8 results in greater speaker errors.

For the last segment (orders greater than 12), we see a smaller drop in the performance as the order is increased. We note that the LP residual contains both modeling and excitation errors. As the LP order increases beyond 10, the residual signal is mainly the excitation error.

It is also interesting to note that residual obtained by 2^{nd} order prediction performs slightly better than the baseline MFCC features in both SDM and MDM cases. Revisiting the performance versus privacy tradeoff from Section 3.7.3, an LP order of 8 seems appropriate for the diarization task, since the first two formants are important for synthesizing an intelligible speech signal (Donovan, 1996). At this prediction order, residual yields a performance of 22.3% on the MDM data while

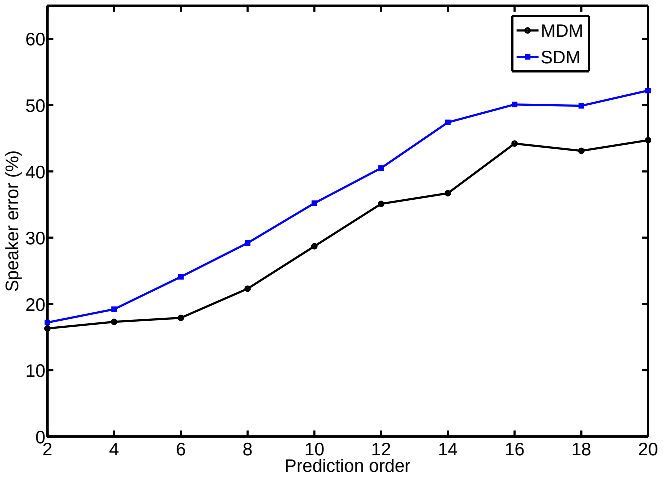

Figure 5.6: Using MFCC representation of LP residual, prediction order vs speaker error is illustrated on MDM and SDM conditions of the development dataset (RT06).

yielding 29.2% on the SDM data.

Combination with Subband and Slope Features

The effect of combining LP residual of 8^{th} order in MFCC representation with slope and subband on MDM data is presented in Figure 5.7. The x-axis denotes the weight assigned to LP residual, while the y-axis denotes the speaker error. We ran experiments varying the weights in steps of 0.05 starting from 0.05 to 0.95. A weight of 1 denotes that LP residual is used without the other features.

It can be observed from the plot that for either slope or subband energies, combining the residual with weights less than 0.45 yields a lower performance than that achieved with LP residual alone. In general, combination with the subband energies yields a slightly better performance over slope at smaller weights. On the other hand, for weights over 0.4, the plot shows that the difference between slope and subband energies may not be significant. For instance, the best combination with spectral slope yields a speaker error of 20.7% at a weight of 0.45, while the best combination with subband energy yields a speaker error of 20.9% at a weight of 0.6.

We note that combining the LP residual with both slope and subband energies yields a consistent gain over combination with either features. The best performance of this combined system is 18.6%

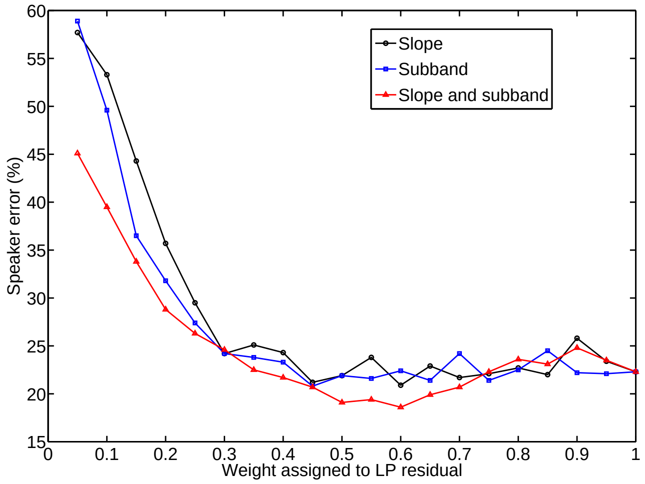

Figure 5.7: Combination of LP residual (MFCC representation) with slope and subband on the development dataset (RT06). X-axis denotes the weight assigned to the LP residual.

at a weight of 0.6. At this configuration, these features yield a promising comparison with the baseline MFCC features (17.1%). It is interesting to note that the diarization system which models the features using Gaussian distributions is suitable for the proposed features as well.

5.5.2 Deep Neural Network Architecture

We now analyze the parameter selection issues associated with the deep neural architecture, namely input domain, bottleneck size, and filtering domain.

The phoneme and the reconstruction MLPs were trained on the TIMIT train dataset. Using these MLPs, filtered squared magnitude vectors, as discussed in Section 5.3.2, were obtained on the MDM development data (RT06 eval). Furthermore, MFCC representation was obtained from these filtered squared magnitude vectors.

Figures 5.8 and 5.9 illustrate the effect of bottleneck size versus speaker error rates on the development data. The input features are squared magnitude and MFCC vectors, respectively. The size of the reconstruction MLP was varied as well. All the other parameters of the phoneme MLP and the reconstruction MLP were unchanged during the experiments.

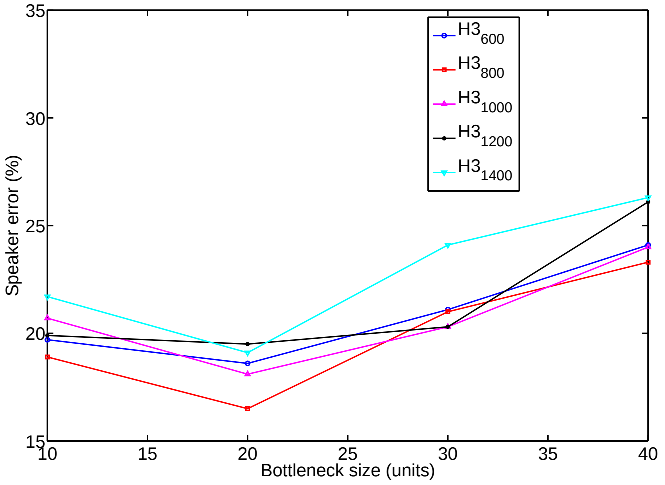

Figure 5.8: Performance of the deep neural network on the development data (RT06). Bottleneck size (B - in terms of number of units) versus speaker error rates (%) for 5 different reconstruction layer sizes (H3). The input features are squared magnitude vectors.

Squared Magnitude Input

For the experiments in Figure 5.8, the input to the phoneme MLPs was 257 dimensional squared magnitude vectors. The output of the reconstruction MLP was 257 dimensional squared magnitude vectors as well. We varied the bottleneck sizes from 10 to 40 in steps of 10. This was repeated for 5 different reconstruction layer sizes from 600 to 1400, in steps of 200. Preliminary experiments indicated 1000 nodes to be a reasonable choice for the first and the third expansion layers of the phoneme MLP - i.e., H1 and H3 in Figure 5.2.

From Figure 5.8, it can be observed that, in general, for all reconstruction layer sizes, a bottleneck layer size of 20 units seems to yield the lowest speaker error rates. When the number of units are higher or lower, the speaker error increases. A similar trend was observed for a 5 layer MLP architecture in (Grezl and Fousek, 2008). We could infer that a bottleneck size of 20 units is sufficient to capture phoneme information using a bottleneck architecture. With a larger bottleneck, some speaker information could be captured.

Furthermore, the “optimal” size of the expansion layer in the reconstruction MLP is around 800 units. In general, for either more or less number of units, we observe an increase in the speaker errors for the other bottleneck sizes. Intuitively, the reconstruction MLP is trying to reconstruct

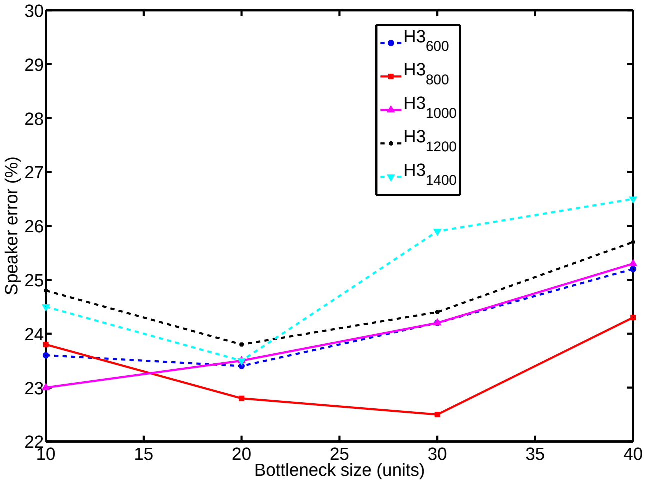

Figure 5.9: With input features as MFCC, performance of the deep neural network features on the development data (RT06). Bottleneck size (B - in terms of number of units) versus speaker error rates ($\%$) for 5 different reconstruction layer sizes (H3).

the input largely with only the phoneme information. Consequently, it is understandable that it requires fewer units (H3) than the first expansion layer (H1) of the phoneme MLP.

We remark that the deep neural network features obtained from the system with a bottleneck size of 20 yields a performance of 16.5% on the MDM development data, which represents a gain of 0.6% over the baseline MFCC features.

MFCC Input

We now examine Figure 5.9, where the input of the phoneme MLP was 19 dimensional MFCC. The output of the reconstruction MLP was 257 dimensional squared magnitude vectors. Bottleneck sizes were varied from 10 to 40 in steps of 10, for 5 different reconstruction layer sizes from 600 to 1400, in steps of 200.

Experiments indicated that 1000 nodes is a reasonable choice for the first and the third layers of the phoneme MLP. Although a bottleneck size of 30 in conjunction with a reconstruction layer size of 800 yields the lowest error, having 20 units for the bottleneck layer seems to be the most reasonable choice. Furthermore, reasonable size for the expansion layer of the reconstruction MLP again appears to be 800 units.

Filtering Domain

We performed studies on MFCC being the output of the reconstruction MLP. Unfortunately, the results were not satisfactory. Since the objective of the chapter was not to optimize all the parameters of the proposed deep neural architecture, but to analyze the feasibility of the architecture itself, we chose not to delve into the details of why MFCC may not be the optimal filtering domain.

Selected Deep Neural Network Architecture

In conclusion of the analysis in this section, we choose the deep neural architecture with log-squared magnitude input (257-dimensional input), 1000 units for the first expansion layer (H1) of the phoneme MLP, 20 units for the bottleneck layer (B), 1000 units for the second expansion layer (H2) of the phoneme MLP, and 800 units for the expansion layer (H3) of reconstruction MLP. The output is a 257-dimensional log-squared magnitude input.

5.6 Diarization Results on RTeval07

Recall that we use RTeval07 as the evaluation dataset. The results of diarization experiments on MDM and SDM conditions are reported followed by results on phoneme recognition. The relationships suggested by feature analysis is then analyzed.

5.6.1 Baseline MFCC

Table 5.2 lists the performance of the baseline diarization system RT07 MDM and SDM evaluation data. The performance of the speech/nonspeech detection system on the RT07 evaluation dataset is 3.7%. On RT07 evaluation data, we observe an even higher performance difference for the MFCC features between the SDM and the MDM, with the actual difference being 4.8%.

Table 5.2. RT07 evaluation data: The first 3 columns list the performance of the speech/nonspeech detection while the next 2 columns report the performance of baseline MFCC features for MDM and SDM.

Evaluation	Miss	FA	sp/nsp	Spkr err (%) MDM	Spkr err (%) SDM
RT07	3.7	0.0	3.7	**6.4**	**11.2**

Table 5.3. RT07 evaluation data: Performance of 8^{th} order LP residual and deep neural network based features. LPR8 denotes LP residual represented using MFCC. SB denotes subband information from 2.5 kHz to 3.5 kHz, while SS denotes spectral slope.

Features	Spkr err (%) MDM	Spkr err (%) SDM
MFCC (baseline)	6.4	11.2
LPR8	12.9	12.0
LPR8 + SB	11.9	11.9
LPR8 + SS	11.3	12.2
LPR8 + SB + SS	11.0	11.5
DNN	14.5	13.9

5.6.2 Comparison with MFCC on RT07 MDM

Table 5.3 lists the diarization results in MDM and SDM conditions. As part of notation, LPR8 denotes 8^{th} order LP residual represented using MFCC, while SB and SS denote subband (2.5 kHz to 3.5 kHz) and spectral slope, respectively. DNN denotes the deep neural network features.

It can be observed that the baseline MFCC yields the best speaker errors on MDM. As a matter of interest, baseline MFCC in combination with Time-Delay Of Arrival (TDOA) features yields a speaker error of 10.9%. The addition of TDOA does not always lead to an improvement, as discussed in Vijayasenan (2010).

LPR8 yields a performance that is 6% below MFCC's, a trend that was observed on the development data. Similarly, combining LPR8 with either SS or SB, yields a gain. This shows that SS and SB have information complementary to LPR8. Combination with both SS and SB yields a gain of nearly 2%; however, the difference with MFCC is still 4.6%.

Table 5.3 shows that DNN yields a performance of 14.5% on MDM. This represents a performance drop of nearly 8% in comparison to baseline MFCC. This result is similar to that of residual features. We shall analyze these errors at the level of each meeting subsequently.

5.6.3 Comparison with MFCC on RT07 SDM

We now focus on the results obtained on the RT07 SDM condition, presented in the third column of Table 5.3.

Consistent with the results on MDM, MFCC still yields the best result. This shows that there is useful speaker information in the first few formants – although higher order formants tend to carry more speaker information Sambur (1975) – that are removed by LP analysis as well as by

DNN. These conclusions are supported by our results for speaker change detection in Parthasarathi *et al.* (2009b), where the addition of energies from a lower subband (1.5 kHz to 2.5 kHz) yielded improvements to residual, although not to the extent of subband (2.5 kHz to 3.5 kHz).

While MFCC does not perform worse than the proposed features on SDM, the change from MDM to SDM results in a smaller difference in speaker error between MFCC and residual features (0.8%). This result could be attributed to LP residual capturing instants of significant excitation, an aspect that has been exploited earlier in Murty *et al.* (2007). Adding either spectral slope or subband information to LPR8 does not yield a gain, however, adding both yields a small gain of 0.5%.

From Table 5.3, it can be seen that DNN yields a performance of 13.9% on the SDM data. This represents a performance drop of 2.7% in comparison with baseline MFCC. It also appears that DNN features are less sensitive to the change from MDM to SDM. We attribute this to reasons similar to that of residual, since Figure 5.3 shows that the DNN approach captures pitch information.

5.6.4 Meetingwise Comparison

Table 5.4 presents a summary statistics of the dataset, with the average length being 43 minutes. The longest meeting is 70 minutes, while the shortest meeting is 25 minutes. In almost all meetings there are 4 speakers, with the exception of NIST-20060216-1347 and VT-20050408-1500, where there are 6 and 5 speakers, respectively.

Table 5.4. Statistics of the RT07 evaluation dataset.

S.No	Meetings	Length minutes	Speakers	Turns
1	CMU-20061115-1030	41	4	758
2	CMU-20061115-1530	29	4	708
3	EDI-20061113-1500	50	4	873
4	EDI-20061114-1500	48	4	557
5	NIST-20051104-1515	70	4	650
6	NIST-20060216-1347	47	6	630
7	VT-20050408-1500	25	5	508
8	VT-20050425-1000	35	4	726

Figure 5.10 compares the speaker errors on MDM and SDM conditions for each meeting. The upper plot shows the comparison on MDM while the lower plot shows it on SDM. The first 8 blocks correspond to the 8 meetings in the evaluation dataset, while the ninth block corresponds to the

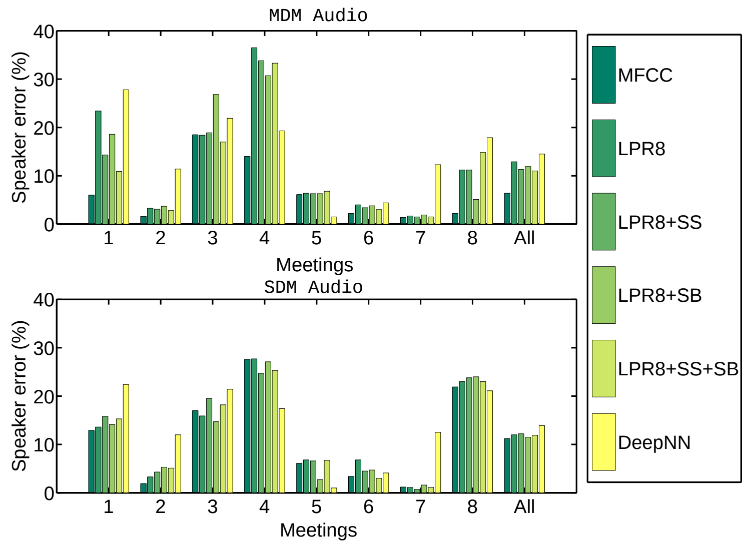

Figure 5.10: Meetingwise analysis of the 9 meetings in the RT07 evaluation dataset. The upper plot shows the comparison on the MDM audio while the lower plot shows the comparison the SDM audio. The meeting numbers correspond to the first column in Table 5.4.

entire dataset.

On MDM, not only does MFCC perform better than residual and DNN features on the whole data, it performs better on most meetings. This supports our analysis in the previous subsection. However, this performance difference diminishes when the average turn length per meeting is longer or when the meetings themselves are longer. Similarly, while the addition of spectral slope and subband information to residual translates to a gain in performance in most meetings; again, this gain is smaller when the average turn length is longer or when the meetings are longer. It appears that in these cases, extra information – in MFCC or in SS and SB – aids speaker discriminability.

On SDM, residual features are comparable to MFCC on most meetings. Furthermore, it is reassuring to observe that the gains, albeit small, due to the addition of SS and SB to LPR8, are more for meetings with shorter turns. These results support our analysis on MDM as well on the whole data. DNN features exhibit similar trends observed on MDM.

5.6.5 Obfuscation Method

In Section 5.3.1, we mentioned another strategy that can be gainfully employed for improving the privacy of audio features. In this section, we present speaker error rates of MFCC and LPR8 features that are randomized with block sizes ($N = 1, 5, 9, 13$) on the RT07 MDM evaluation dataset in Table 5.5. In the table, "Randx" is used to denote randomization with block size x frames. We note that randomizing the MFCC features with various block sizes does not change the performance significantly ($\leq 1\%$). Similarly, the performance of the LP residual remains unaffected by local temporal randomization. This is largely due to the fact that speaker information spans long temporal spans, and such an obfuscation procedure does not affect the speaker change points much.

Table 5.5. Effect of randomization on MFCC and LPR8 on the RT07 MDM dataset. Randx is used to denote randomization with block size of x frames.

Feature	LPR8 (%)	MFCC
	Spkr err	Spkr err
Rand5	13.4	6.7
Rand9	13.8	7.1
Rand13	13.7	6.8

5.7 Privacy Analysis

So far we have investigated LP residual and deep neural network based features for speaker diarization. We now proceed to make an analysis of the privacy aspects.

In Chapter 3, we explored phoneme recognition as a means to assess privacy. In addition to this, in this chapter, we explore another method to analyze the notion of privacy as the linguistic information. Specifically, we explore the human speech recognition rates (HSR) on synthesized speech obtained from privacy-sensitive and MFCC features.

5.7.1 Analysis using Human Speech Recognition

In the field of HSR, one aspect of the listening test is whether the vocabulary is open set or closed set. Another aspect of these studies is whether one tests on individual units such as nonsense syllables or on fully-formed sentences. Furthermore, fully-formed sentences could be semantically meaningful sentences such as conversations, news, phonetically confusable sentences, or semantically unpredictable sentences

In this study, we used open set, semantically unpredictable sentences (SUS) (Benoit *et al.*, 1996). This is done so that the test evaluates only the acoustic aspect of intelligibility instead of the cognitive aspect of prediction. SUS are usually constructed from simple grammatical templates.

HSR Setup

For our experiments, we used the 20 SUS from the EMIME bilingual database (Wester, 2010), with a vocabulary size of 88 words. The list of sentences is given in Table 5.6. In this database, there are 7 female and 7 male native english speakers with different accents. We chose one female and one male speaker, resulting in 10 sentences being spoken by female and 10 being spoken by male speakers. The speech from the close talking microphone, sampled at 22 kHz, was downsampled to 16 kHz.

We generated the following features from this audio: (a) baseline MFCC features; (b) MFCC representation of 8^{th} order LP residual; and (c) MFCC representation of deep neural network features. From these MFCC representations, noise-excited reconstructions were obtained. Reconstruction was done using the RASTAMAT library: `http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/`. It has to be noted that for the residual represented using MFCC, such a reconstruction yields a reconstructed residual signal. Since the reconstruction was noise-excited, they sound whispered.

Upon reconstruction, we now have audio from the 3 sets of features for each of the 20 sentences. Since our pool of listeners were mostly non-native in English, we added the raw waveform as the 4^{th} set of audio (or 4^{th} *system*) for the 20 sentences. This is done to estimate the upper bound of performance that can be achieved by non-native listeners.

Because we expected few listeners (and eventually had 27), in the tradeoff between reasonable estimates of intelligibility versus repeating each sentence as few times as possible, we chose the following strategy: we divided the 80 utterances (20 sentences × 4 systems) into 2 groups of 40 each. Each group of 40 utterances were obtained with a Latin square design to maximize coverage of the four systems and the 20 sentences. In order that listeners do not get used to a predetermined sequence of audio from the 4 systems, we randomized the sequences in both groups. Each listener

Table 5.6. 20 semantically unpredictable sentences in the dataset.

No.	Sentence
1	The dust leaned through the broad hat.
2	The task joined the staff that coped.
3	The pure word cleaned the mind.
4	When does the flow guide the blue front?
5	Use the length or the export.
6	The youth knelt with the fresh state.
7	The road dared the growth that slipped.
8	The large wine blamed the store.
9	How does the thing cut the true wall?
10	Bear the truth and the pool.
11	The foot gazed under the dead spring.
12	The suspect mixed the pain that crept.
13	The nice block paid the blood.
14	Why does the jazz hit the brown bar?
15	Bite the book and the stress.
16	The health went down the dark square.
17	The dog built the wife that walked.
18	The good man marked the tree.
19	Where does the post need the poor race?
20	Export the son or the firm.

was assigned to one of the two groups and she/he listened to 40 utterances with 10 utterance from each system. Each listener, therefore, listened to each sentence twice. This, of course, has the effect of the listener performing better on the test – which in our case yields a lower bound on the privacy estimate.

A web-based application was setup so that listeners could listen using their headphones or speakers. After listening, they had to type-in the sentences they heard. They could complete the task in multiple sessions. Listeners were asked to restrict the number of times they could listen to an utterance to a maximum of 5 times. If an utterance was not intelligible after that, they were asked to type “Not intelligible”. Out of the 27 listeners who did the test, one was a native English listener.

HSR Experiments

Before scoring, we preprocessed the listeners’ typed-in responses. This was done to ensure that spelling mistakes or punctuation marks do not show up as errors in intelligibility. For example, some listeners used ellipsis or “?” to indicate words they missed. These were removed from the responses. We used the HResults tool (Young *et al.*, 2000) to score the number of words correctly recognized. This is the ratio of number of correct words to the total number of words.

The results of scoring the features are listed in Table 5.7. In addition, we also obtained an ordering of listeners according to the percentage of words correctly recognized. In Table 5.7, the two rows correspond to the performance of the 4 systems scored over all the listeners, or scored only over the top 10 best performing listeners over all conditions. The four columns indicate performance corresponding to the 4 systems: (a) raw waveform; (b) reconstruction from MFCC; (c) reconstruction from MFCC representation of 8^{th} order LP residual; and (d) reconstruction from MFCC representation of deep neural network features.

Table 5.7. HSR performance, in terms of word accuracy, of the 4 systems over all the listeners or over the top 10 best performing listeners. The four columns indicating performance correspond to raw waveform, reconstruction from MFCC, from MFCC representation of 8^{th} order LP residual, and from MFCC representation of DeepNN features, respectively.

	Wav	MFCC	LPR_8^{MFCC}	$DeepNN^{MFCC}$
Total	85.2	71.3	***13.7***	**6.8**
Top-10	91.8	79.4	***28.9***	**16.9**

It can be seen that for both sets of listeners (total, and top-10), listening to the raw waveform yielded the best performance. Reconstruction from MFCC also yielded very good intelligibility, i.e., around 71% for all the listeners and around 79% intelligibility for the top-10 listeners. In general, listening to speech reconstructed from the MFCC representation of 8^{th} order LP residual appears much less intelligible, with around 50 % to 60 % drop in intelligibility. This could partially be due to the loss of the first formant, which carries more linguistic information (Donovan, 1996). In addition, there is a further loss in information from LP residual by representing it using MFCC. Deep neural network based features yield the lowest intelligibility, yielding around 7% intelligibility over all listeners and around 17% over the top-10 listeners.

Furthermore, since listeners listen to each sentence twice, some listeners reported that this led to them performing better on systems having lower intelligibility (having already listened to a cleaner version before). On the other hand, the two sequences corresponding to the utterances for each group were randomized and therefore there is no systematic bias towards privacy-sensitive or the non privacy-sensitive systems.

5.7.2 Analysis using Automatic Phoneme Recognition

This section is similar in spirit to Section 3.7 from Chapter 3. We perform phoneme recognition experiments to assess the privacy-sensitive (LP residual and deep neural network) and the MFCC features. While the trends for LP residual from both sections are similar, the actual performance figures are slightly different due to the difference in representation between MFCC and MFPLP. This thesis, however, does not focus on these differences.

Phoneme Recognition Setup

We recall the phoneme recognition setup in this subsection. In our experiments, phoneme recognition studies were performed on TIMIT database. Experiments were conducted excluding the 'sa' dialect sentences. The training data consists of 3000 utterances from 375 speakers, cross-validation data consists of 696 utterances from 87 speakers, and the test data set consists of 1344 utterances from 168 speakers. The phoneme set corresponds to the standard set of 39 units (Lee and Hon, 1989).

Features are mean/variance normalized across the training data set. A three layered MLP is used to estimate the phoneme posterior probabilities. The MLP consists of 1000 hidden units, and 39 output units with softmax nonlinearity, representing the phoneme classes. The input layer uses a temporal context of 9 frames on the features generated at a frame rate of 100 Hz. For all the features studied (baseline MFCC, LP residual with MFCC representation, deep neural network features with MFCC representation), the input to the MLP was 13-dimensional MFCC with delta and acceleration coefficients. The MLP is trained using standard back propagation algorithm by minimizing the cross entropy error criterion. The phoneme recognition experiments are performed using the hybrid HMM/MLP system reported in (Bourlard and Morgan, 1994). The phoneme sequence is decoded using the Viterbi algorithm, where each phoneme is represented by a left-to-right, 3-state HMM, enforcing a minimum duration of 30 ms. The output distribution in each of the three states is the same, and is derived from the output of the MLP.

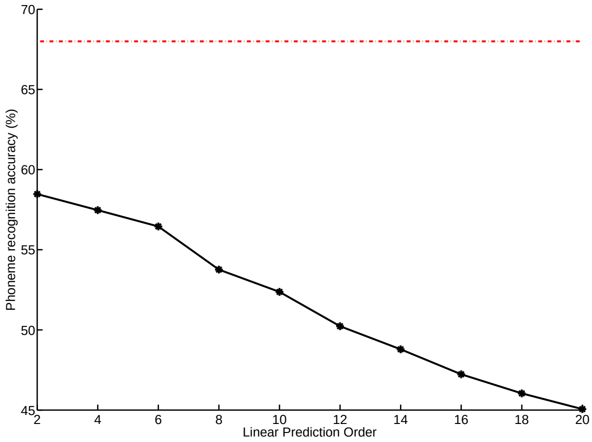

Figure 5.11: Phoneme recognition accuracy for the residual based features various LP orders on TIMIT. The x-axis shows the LP order while the y-axis shows the phoneme accuracy in (%).

Phoneme Recognition Experiments

Figure 5.11 plots the recognition accuracies with respect to increasing LP orders using the phoneme recognition system. It can be observed that as the LP order increases, the recognition accuracies drop. We note that an increase in LP order by 2 can allow an extra complex conjugate pole pair to be modeled, possibly modeling an extra formant. Since lower order formants generally carry more linguistic information, one could expect the performance to drop when the LP order is increased.

From Figure 5.11, we observe that the LP residual with a prediction order of 8, yields around 15% lower phoneme recognition accuracy in comparison with the MFCC features. We remark that the phoneme recognition experiments using simple features proposed in (Parthasarathi *et al.*, 2010), namely, spectral flatness, energy, zero-crossing rate, and kurtosis (*SEZK*)and the features proposed in (Wyatt *et al.*, 2007a), namely, autocorrelation and relative-spectral entropy (*AH*), with delta and acceleration coefficients, and with a 9 frame context, yielded accuracies of 40.8% and 31.2% respectively. The performance of an 8^{th} order LP residual lies between that of the simple features and the MFCC (68.2%).

Phoneme recognition experiments using the MFCC representation of deep neural network features yielded 48.7%. This recognition accuracy is much lower than that of 8^{th} order LP residual.

We then performed recognition experiments for the obfuscation method on 8^{th} order LP residual.

We note here that randomization can be performed for (a) only test data; or (b) both train and test data with different seeds. The difference between the two stems from the fact that in the second case, the MLP has been trained with noisy targets. While randomized training (29.3%) improves the performance marginally over clean training (28.2%), we still observed a substantial drop in phoneme recognition performance over residual itself. Although our HSR experiments in the previous section showed that reconstructing speech from MFCC representation of 8^{th} order LP residual is unintelligible, this result suggests that randomization can be used to enforce further privacy.

5.8 Summary and Conclusion

In this chapter we presented two different approaches to privacy-sensitive audio features for robust speaker diarization, namely, LP residual based and deep neural network based. We systematically investigated both sets of features for speaker diarization in single and multiple distant microphone conditions. The SDM scenario, however, is more relevant to a portable audio recorder scenario. The notion of audio privacy is interpreted as linguistic privacy. Our investigations are summarized in Table 5.8 and they suggest the following.

Table 5.8. Summary of this chapter in terms of (a) Diarization performance on SDM data using speaker errors (%) and (b) privacy assessment – using automatic phoneme recognition accuracy (%) and human word recognition accuracy from synthesized speech (%).

Features	SND Results	Privacy assessment	
	Speaker errors (%)	Phoneme accuracy (%)	Intelligibility (%)
Residual features	11.5	52.9	28.9
Randomized residual features	13.7	28.2	-
Deep neural network features	13.9	48.7	16.9
State-of-the-art	11.2	68.2	79.4

LP Residual

We studied two different strategies to represent the LP residual, with the MFCC representation of the residual yielding superior performances for all prediction orders. Additionally, we explored the combination of residual with subband information from 2.5 kHz to 3.5 kHz and spectral slope. Although residual features performed slightly less than the conventional MFCC features, we observed that residual features are less affected by the change from MDM to SDM scenarios. Furthermore, residual features proved to be more privacy-sensitive than MFCC features in terms of lower intelligibility and phoneme recognition accuracy.

Deep Neural Network

We utilized a greedy, layer-by-layer trained deep neural network for representing the phoneme information in the short-term spectrum of the signal. A second MLP was utilized to reconstruct the spectrum. The reconstructed spectrum was used as a phoneme filter. In terms of diarization performance, this approach performed slightly worse than the LP residual based approach. However, these features proved to be more privacy-sensitive then residual features. Future work on this approach will investigate improvements such as training the deep MLP on meeting data.

Putting Privacy and Diarization Together

We attempted to quantify the abstract notion of privacy in audio through phoneme recognition and intelligibility studies. On the one hand, standard spectral features such as MFCC yielded, not surprisingly, good linguistic reconstruction. Proposed approaches to privacy-sensitive audio feature extraction yielded substantially lower linguistic performance compared to the MFCC features.

While the diarization performance of the LP residual features are similar to the baseline MFCC features, the performance of the deep neural network based features were about 2% lower than MFCC features. However, the effect of a 2% drop in diarization performance on socially relevant tasks such as dominance estimation have been shown to be minimal, if any (Hung *et al.*, 2011).

Chapter 6

Outdoor Conditions

For the most part, the first steps to the nonverbal analysis of face-to-face social interactions using audio data consists of performing speech/nonspeech detection (SND) and speaker diarization (Gatica-Perez, 2006, 2009). This is usually followed by modeling nonverbal features including prosody and other cues like speaking time, speaker turns, and speaker overlaps. To this end, thus far we have studied privacy-sensitive features for SND and diarization and have benchmarked these features against established features such as MFCC, in a variety of conditions, on standard datasets such as meeting room audio. Testing our features for SND and diarization in outdoor conditions would culminate the study of the speech processing aspects of this problem.

One of the problems, however, is the lack of datasets to study this setting. Almost all of the few available datasets cannot either be redistributed, or are not available as raw audio due to privacy concerns. For example, the data collected using the original sociometer (Choudhury, 2004) is not shareable, as for instance is the data collected by Ellis and Lee (2006). One exception is Conversational Speech in Noisy Environments (COSINE) (Stupakov *et al.*, 2009), an outdoor conversational dataset collected by subjects walking around a university campus.

On this dataset, we begin with the problem of SND; and this chapter not only compares the proposed features (simple features and LP residual) against the MFPLP features under these conditions, but also investigates if a set of trained indoor SND models suffice for speech detection in outdoor conditions. This problem of data and model mismatch is particularly relevant with real-world data collection. Consequently, we explore, in addition to the traditional combination methods (of classifiers/detectors) such as *averaging* and *inverse entropy*, a maximum likelihood (ML) combination scheme which promises an *oracle-like* combination – of which more will be said subsequently. We find a slight improvement with the ML scheme over averaging and inverse entropy, although there is still a substantial performance gap with the models trained on outdoor conditions. Nevertheless, our work on SND in this chapter confirms the conclusions in Chapter 3 that models trained on privacy-sensitive features can yield a performance similar to models trained on state-of-the-art features.

Our results from Chapter 5 show that features derived from the eighth order LP residual and deep neural networks yield a more privacy-sensitive representation, while yielding "near state-of-the-art" performance for diarization. Out of these features, our results had shown that the LP residual yielded a more promising performance for diarization, albeit with a lower privacy. This chapter proceeds to investigate speaker diarization in outdoor conditions using the LP residual and benchmarks it against MFCC features. The effect of SND on the diarization performance is then briefly analyzed on the outdoor conditions. We found that using the privacy-sensitive SND led to a small drop in performance.

The rest of this chapter is organized in three sections. Section 6.1 begins with a brief description of the outdoor dataset. This is followed by two sections, on SND and diarization, respectively. In Section 6.2, we review traditional classifier combination schemes for SND followed by an explanation of the proposed ML method for combining indoor detectors. SND results are presented next. In Section 6.3, we first present the diarization results of the LP residual features on the outdoor dataset. Section 6.4 presents the conclusions of this chapter. The material presented in this chapter is presented here for the first time.

6.1 Outdoor Dataset

Conversational Speech in Noisy Environments (COSINE) readily provides us with an outdoor dataset that was collected by subjects engaged in spontaneous conversations, wearing multiple microphones (Stupakov *et al.*, 2009, 2011). These conversations were recorded by participants walking around the University of Washington, Seattle campus. Privacy-sensitive portions of the audio, corresponding to the last name of a participant or an account number or any segments that the participants wished to be deleted, were set to zero in the captured audio signal. Each participant wore 7 microphones; these consisted of four array microphones worn around the waist, one close-talking microphone, one shoulder microphone, and one throat microphone.

COSINE contains both transcribed and untranscribed sessions, the duration of each of which is between 45 minutes to 1.5 hours. Ten sessions are fully transcribed and it amounts to a total of 36.5 hours of transcribed audio - with speech and nonspeech being in the ratio 2.7 to 1. The audio has transcriptions in terms of speakers, speech, and the privacy-sensitive segments per speaker. It has to be noted that each session is captured multiple times due to each participant wearing multiple microphones. The total untranscribed audio amounts to nearly 145 hours.

For our experiments, we used the transcribed sessions. After some listening, we found the audio from the array microphones to be more intelligible for far-field speakers. Consequently, the audio from one of the four array microphone channels (from one person per session) was selected and downsampled from 44.1 kHz to 16 kHz. This yields around 10 hours of audio, with 15 male and 22 female speakers. However, transcriptions for one of the speakers was missing in session 4; and consequently, this session was not considered in our experiments. By merging speaker-level transcriptions, and by accounting for the wearer of the array microphone, close-talking and far-field speakers were identified, and a groundtruth to that end was created. Training, development, and testing data were created by partitioning the data into 4, 1, and 4 sessions, respectively. Privacy-sensitive segments of each of the sessions were not scored because there were no real observations. As a fraction of the overall data, the privacy-sensitive manual deletions were negligible.

6.2 Speech/nonspeech Detection

The objective of this section is three-fold: (a) assess the performance of indoor models on the outdoor conditions; (b) assess the fusion methods by combining the indoor models on the outdoor conditions; and (c) assess the performance of the outdoor models on the outdoor conditions. We begin this section with a brief description of the combination methods – in particular, three methods including the proposed ML method is discussed. This is followed by a description of the experimental setup and the results.

6.2.1 Classifier Combination

One of the objectives of classifier combination (Kittler *et al.*, 1998) is to exploit the complementary information between the classifiers. Combination techniques typically combine either the decisions made by the individual classifiers or assign a weight to each classifier's evidence. These weights can be either estimated statically (on cross-validation data) or dynamically. In this chapter, we consider three weight allocation strategies: (a) dynamic weighting using inverse entropy (Kittler *et al.*, 1998); (b) static weighting using equal weights/averaging (Kittler *et al.*, 1998); and (c) static weighting using maximum likelihood (proposed method).

Inverse Entropy

Inverse entropy-based classifier combination has been used in ASR studies (Misra *et al.*, 2003). Let $c \in \{s, n\}$ denote the speech/nonspeech classes and let x_t^k denote a feature k at a time t. $P(c|x_t^k; \theta_k)$ denotes the posterior probability estimate obtained from the MLP classifier trained on a feature k, and θ_k denotes the MLP model parameters for a feature k. Inverse entropy based combination assigns larger weights to classifiers that are more confident as measured by the entropy (h_k) of its posterior probabilities (Misra *et al.*, 2003). The weights for the k^{th} classifier are then estimated as:

$$w_k = \frac{\frac{1}{h_k}}{\sum_j \frac{1}{h_j}} \tag{6.1}$$

The combined evidence using all the features X_t^k:

$$P(c = i|X_t^k) = \sum_k w_k \cdot P(c = i|x_t^k; \theta_k) \quad \forall i \in \{s, n\} \tag{6.2}$$

Averaging

In this technique Kittler *et al.* (1998), all the classifiers are assigned equal weights, i.e., $w_k = \frac{1}{N}$. The output evidence is combined using equation 6.2.

Maximum Likelihood Combination

We propose a technique based on the assumption that the output of the detectors trained on indoor conditions can be viewed as noisy versions of an underlying groundtruth; furthermore, we assume that this "noise" is Gaussian. We shall now discuss this method of combining the detectors in this section.

Let y_i^j be random variables denoting the output of j detectors for each frame (i) of speech. We shall assume a simplistic model where we consider these outputs are perturbed around a mean y_i, with a variance σ_j^2. This is equivalent to the assumption that the errors committed by each detector around the true decision (y_i) are Gaussian with zero mean and variance (σ_j^2). More formally,

$$\begin{aligned} \epsilon_j &\sim \mathcal{N}(0, \sigma_j^2) \\ y_i^j &= y_i + \epsilon_j \\ p(y_i^j|y_i, \sigma_j) &\sim \mathcal{N}(y_i^j, \sigma_j^2) \end{aligned} \tag{6.3}$$

Let $\mathcal{D}$ denote the data, and let Θ denote the set of parameters y_i, σ_j, where $i \in 1 \cdots N$, and $j \in 1 \cdots R$. Then the likelihood function can be written as,

$$\mathcal{L}(\Theta) = p(\mathcal{D}|\Theta) = \prod_{i=1}^{N}\prod_{j=1}^{R} p(y_i^j|y_i, \sigma_j) = \prod_{i=1}^{N}\prod_{j=1}^{R} \mathcal{N}(y_i^j, \sigma_j^2)$$

The maximum likelihood estimate of the parameters can be formulated as,

$$\begin{aligned} \widehat{\Theta}_{ML} &= \arg\max_{\Theta} \big(\ln p(\mathcal{D}|\Theta) \big) \\ &= \arg \max_{\{y_i, \sigma_j\}} \sum_{i=1}^{N}\sum_{j=1}^{R} \ln \Big(\frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(y_i^j - y_i)^2}{2\sigma_j^2}} \Big) \\ &= \arg \max_{\{y_i, \sigma_j\}} -\sum_{i=1}^{N}\sum_{j=1}^{R} \Big(\frac{1}{2} \ln 2\pi\sigma_j^2 + \frac{(y_i^j - y_i)^2}{2\sigma_j^2} \Big) \end{aligned} \tag{6.4}$$

Computing the derivatives of $\mathcal{L}$ with respect to the parameter set $\{y_i, \sigma_j\}$ and solving, we get

$$\widehat{\sigma_j}_{ML} = \frac{1}{N} \sum_{i=1}^{N} (y_i^j - y_i)^2 \tag{6.5}$$

$$\widehat{y_i}_{ML} = \frac{\sum_{j=1}^{R} \frac{y_i^j}{\sigma_j^2}}{\sum_{j=1}^{R} \frac{1}{\sigma_j^2}} \tag{6.6}$$

As the parameters $\widehat{\sigma_j}$ and $\widehat{y_i}$ are coupled together we iterate these two steps till convergence.

The result is intuitive as it estimates the groundtruth as a weighted combination of the output of the detectors. The weights themselves are estimated as the inverse of the variance of the detectors. Furthermore, similar results have been obtained in multimodal combinations, for example in Ernst and Buelthoff (2004).

6.2.2 Experimental Setup

This section discusses the experimental setup to assess the three objectives listed at the beginning of Section 6.2. We begin with a description of the indoor and the outdoor SND models. This is followed by the evaluation measure used to assess the models/systems.

Indoor models

Indoor SND models based on four sets of features are evaluated on the outdoor dataset. For each of the following features, a corresponding indoor model/system (trained MLPs) exists:

- Two systems derived from simple features: AH and EZK. Please refer to Chapter 3 for details.
- One system from the MFPLP features derived from 8th order LP residual;
- One system from the MFPLP features.

Furthermore, as described in Section 3.3, the models are trained using these features on three different indoor datasets, namely, AMI, ICSI, and NIST. There are, therefore, a total of 12 indoor SND systems. These 12 indoor detectors/systems are summarized below:

- On AMI: 3 privacy-sensitive MLP systems using AH, EZK, and LPR8 as features respectively, and one state-of-the art MLP system using MFPLP.
- On ICSI: 3 privacy-sensitive MLP systems using AH, EZK, and LPR8 as features respectively, and one state-of-the art MLP system using MFPLP.
- On NIST: 3 privacy-sensitive MLP systems using AH, EZK, and LPR8 as features respectively, and one state-of-the art MLP system using MFPLP.

Recall from Chapter 3 that combining spectral flatness (S) with Energy (E), Zero crossing rate (Z), and Kurtosis (K) did not yield gains in performance. This trend was observed in this chapter as well. We therefore use the system (EZK) instead of the system (SEZK).

Combination of indoor models

Three combination methods described earlier, namely averaging, inverse entropy, and the ML method are studied using the nine privacy-sensitive indoor speech detectors. The term *oracle* is specifically used for the *best combinations* (on the COSINE evaluation data) of the 9 privacy-sensitive detectors (using averaging or inverse entropy). This is achieved by manually removing detectors that affect the performance the most. Consequently, we have an “oracle averaging” system and an “oracle inverse entropy” system.

Outdoor models

As a reference, the four sets of features (3 privacy-sensitive using AH, EZK, and LPR8 respectively, and one state-of-the art using MFPLP) are trained on the COSINE data. We therefore have four outdoor models.

Evaluation Measure

The scoring system is identical to that used in Chapter 3. Recall that we plot receiver operating characteristic (ROC) curves for (a) close-talking speech and nonspeech; (b) far-field speech and nonspeech; and (c) all speech and nonspeech. The area under the ROC (AROC) is then computed for the ROCs; this is a measure that is agnostic to changes in the prior distribution of the classes.

6.2.3 SND Results

There are three sets of results: (a) results of the speech detectors trained on the indoor conditions, which are are presented first; (b) this is followed by the results obtained by combining these indoor models; and (c) the last set of results correspond to those obtained by SNDs trained on the COSINE data.

Indoor Detectors

Table 6.1 presents the performance on close-talking and far-field COSINE data by SND trained on indoor conditions. These results are grouped into those that are trained on AMI, ICSI, and NIST.

Table 6.1. Performance on close-talking and far-field COSINE data by speech detectors trained on indoor conditions. The combined performance (over close-talking and far-field audio) of LP residual and MFPLP based detectors are marked with italics and bold respectively.

SNo	Features	Close (%)	Farfield (%)	Total (%)
Detectors trained on AMI				
1	AH	59.4	56.7	57.2
2	EZK	78.5	67.0	69.0
3	LPR8	79.3	69.7	*72.1*
4	MFPLP	82.6	70.5	**73.4**
Detectors trained on ICSI				
5	AH	57.8	56.0	56.3
6	EZK	79.4	69.0	71.2
7	LPR8	82.0	72.6	*75.2*
8	MFPLP	82.6	72.5	**74.9**
Detectors trained on NIST				
9	AH	55.8	56.1	55.8
10	EZK	67.2	61.1	61.8
11	LPR8	66.8	61.7	*63.1*
12	MFPLP	64.1	60.3	**61.3**

The performance of the LP residual and MFPLP based detectors are marked with italics and bold respectively.

In general, it can be observed that the performance of all the systems on the close-talking microphone is better than the performance on the far-field microphone. This is a reasonable result and is also consistent with the results that were obtained in Chapter 3. Furthermore, it can be observed that the performance of the detectors trained on ICSI yields the best performance. This could be due to the fact that among the three datasets, ICSI was the largest training set (50 hours); while NIST and AMI had 10 and 15 hours of training data respectively. While the gap between the performance of the systems trained on ICSI and AMI datasets is smaller (about 2%), the gap between the systems trained on AMI and NIST is quite substantial (about 12 %). Clearly, some justification for this can be drawn from the results obtained from Section 3.6, where it can be seen that training on NIST yields a distinctly lower performance on both ICSI and AMI.

Among the features, MFPLP and LP residual based features achieve the best performance. Furthermore, the difference in performance between the two systems appears close, with the actual outcome as to which system performs better appears to be database dependent. This is consistent with the results that we obtained in Chapter 3. Simple features perform worse than either MFPLP or LP residual features. On all the three datasets, detectors based on EZK yield around 3 – 4 %

lower than MFPLP or LP residual features, a trend similar to that obtained in Section 3.6. On the other hand, AH features yield a substantially lower performance compared to EZK features. While Table 3.4 shows that AH could perform worse than EZK in far-field (and mismatched) conditions, the large drop that we obtain in Table 6.1 requires further analysis.

Combinations of Indoor Detectors

Table 6.2 presents the performance obtained by combining indoor SND models. Similar to the previous section, we group these results into close-talking, far-field, and all conditions. As part of additional notation, the subscripts AI or AIN denote a combination of the systems trained on {AMI and ICSI} or {AMI, ICSI, and NIST}.

The first and the second rows present the results of combinations due to averaging and the inverse entropy denoted by C_m and C_e respectively. For the results in these two rows, we have taken 9 detectors and have combined the results with either method. These 9 detectors are due to systems trained with each of the 3 features (LPR8, EZK, AH) and on three datasets (AMI, ICSI, and NIST). We observe that these combinations do not yield an improvement in performance similar to those obtained in Chapter 3. This could be due to the poor performance of some of the indoor models. To see this, we observe that an averaging or inverse entropy combination of the four-best systems (LPR8 and EZK trained on AMI and ICSI) yield a performance of 77.0 %, which is the best we have obtained so far using indoor detectors. We call these systems the oracle systems. In all these cases, there is no significant difference in performance between the averaging and the inverse entropy based methods – a trend that is consistent with the trends in Chapter 3.

The last row in Table 6.2 presents the results of the ML combination using the 9 systems. It can be observed that the ML combination yields a performance that is close to that of the oracle methods, which is an intuitively satisfying result. Furthermore, the performance of the ML combination is around 2 to 3 % higher than that of averaging and inverse entropy combinations. It could be interesting to define oracle in an alternative fashion – as a trained linear classifier whose inputs are the indoor detector outputs.

On the other hand, one of the issues that we observed with the ML method is that a reduction in the number of detectors affects its performance. For example, combining the four best systems using the ML method does not yield as good a performance (71.1 %) as averaging or inverse entropy. We

Table 6.2. Performance of combinations of indoor SND detectors on close-talking and far-field COSINE data. As part of additional notation, the subscript$_{AIN}$ or $_{AI}$ denote a combination of the systems trained on AMI, ICSI, and NIST or AMI and ICSI. Combinations of detectors achieved by averaging, inverse entropy, and ML schemes are denoted by C_m, C_e, and C_{ml} respectively.

SNo	Systems	Close (%)	Farfield (%)	Total (%)
1	$C_m(LPR8, EZK, AH)_{AIN}$	81.2	71.5	73.6
2	$C_e(LPR8, EZK, AH)_{AIN}$	82.9	72.3	74.8
3	$C_m(LPR8, EZK)_{AI}$ - Oracle	86.6	74.2	*77.1*
4	$C_e(LPR8, EZK)_{AI}$ - Oracle	86.4	74.0	*77.0*
5	$C_{ml}(LPR8, EZK, AH)_{AIN}$	85.3	74.0	**76.6**

hypothesize that, since the estimation of the means affects the estimation of the variance and vice versa, a poor estimation of the means (due to one dominant, "bad" detector) has an adverse effect on the estimation of the weights of the detectors. Another factor could be that having fewer detectors might result in the distribution of the errors (computed as a difference between the detector output and the weighted mean) made by the dominant detector to be less Gaussian that it should be, thereby violating the ML-model assumptions. However, further analysis is needed to ascertain this.

Detectors Trained on Outdoor Conditions

Table 6.3 presents the performance obtained by training the SND on the outdoor data. Similar to the previous sections, performance on the close-talking, far-field, and both the conditions are listed.

The first three rows present the results of the detectors trained on simple features. It can be observed that even over the SND trained with EZK on the ICSI dataset (refer to Table 6.1), a gain of at least 2 to 3 % is obtained. For a feature such as AH, a much more substantial gain is observed.

LP residual features are listed next. Compared to LPR8 features trained on ICSI dataset, a gain of nearly 7 % is obtained. The next four rows present the results obtained by combining the simple features with the LPR8 features. It can be seen that these combinations yield, in some case, an improvement. With this gain, the privacy-sensitive features are comparable to the MFPLP features. On the other hand, the ML method once again does not perform well when we combined the systems due to the three features: LPR8, EZK, and AH. This appears to support our hypothesis that the ML method needs more detectors to obtain reasonable estimates of the groundtruth.

On this dataset, the inverse entropy based combination method shows some gain over the averaging method. In contrast, in the meeting datasets, not only was the performance difference

Table 6.3. Performance of SND detectors trained on COSINE data, and tested on close-talking and far-field COSINE data. Combinations of detectors achieved by averaging and inverse entropy schemes are denoted by C_m and C_e respectively. Bold and italics refer to the best performances achieved by the residual features and the MFPLP features, respectively.

SNo	Features	Close (%)	Farfield (%)	Total (%)
1	AH	80.0	70.0	72.5
2	EZK	81.9	71.8	74.3
3	LPR8	87.6	80.9	**82.9**
4	$C_m(LPR8, EZK, AH)$	87.3	79.9	82.1
5	$C_e(LPR8, EZK, AH)$	88.9	81.2	**83.6**
6	$C_m(LPR8, EZK)$	86.4	78.8	80.9
7	$C_e(LPR8, EZK)$	88.5	81.0	83.2
8	MFPLP	88.4	82.2	*84.0*

between the two methods lesser, but also the averaging methods appeared slightly better. In general, the averaging method is better if the errors made by the detectors are uncorrelated; while the inverse entropy method would yield better results if the most “confident” classifier is also the least error prone. From these statements, it appears that on this dataset, the more confident classifiers are, the less error prone they are too.

6.3 Speaker Diarization

This section reports preliminary speaker diarization experiments performed on the COSINE dataset. For the experiments, we used the audio from the four test sessions, details of which were reported in Section 6.1. Recall that this audio comes from one channel of a four-channel array mic, and that it was originally sampled at 44.1 kHz, which is then downsampled to 16 kHz.

6.3.1 Experimental Setup

Speaker diarization was performed using the ICSI diarization system. For details on this system, please refer to Section 5.4. The baseline diarization system needs a speech/nonspeech segmentation as discussed in Section 6.2. Three sets of speech/nonspeech segmentation were created: (a) from the groundtruth obtained by merging speaker-level word transcriptions; (b) from the output of the detector using MFPLP features, trained on COSINE train data; (c) from the output of the detector using LP residual features, trained on COSINE train data. The groundtruth for evaluating the speaker diarization is created once again by exploiting the speaker-level transcriptions provided by COSINE. To this end, we converted the files from the TextGrid format to the RTTM format. Sessions (7,8, and 9) has 4 speakers, while session 10 had only 3 speakers. The total data amounts to nearly 4.5 hours of audio.

6.3.2 Diarization Results

For the diarization system using SND groundtruth in conjunction with the 19 dimensional MFCC features, we obtained a speaker error of 33.2 %. Clearly this is a much higher speaker error rate than what we observed even on the indoor single channel distant microphone data. However, we are hesitant to claim it as a poor result, as this is one of the first diarization experiments on the outdoor audio.

When SND groundtruth was used in conjunction with LP residual features – 8th order LP residual along with spectral slope and subband energies – speaker error increased by around 3%. By combining SND segmentation obtained by privacy-sensitive or non privacy-sensitive systems with the corresponding features (LP residual features or MFCC features obtained from raw audio, respectively) for diarization, the speaker error rate increased further.

We could attempt to improve the performance of these systems in a few ways. One approach is to tune the parameters of the diarization systems. For the MFCC based system, these include the number of Gaussians, initial number of clusters, and the minimum duration. In addition to these, for the residual based system, the weight to combine the likelihoods of the feature streams could be tuned. Another approach is to understand the effect of background noise. Since this is an outdoor data, the background noise is clearly non stationary, and perhaps the diarization system clusters the backgrounds rather than the speakers. For this, we could explore techniques such as RASTA filtering (Hermansky and Morgan, 1994).

6.4 Conclusions and Future Work

In this chapter of the thesis, we studied the performance of our SND and diarization methods on outdoor audio – a condition which we believe to be pervasive in the capture and analysis of real-world audio. We explored the suitability of using SND models trained on indoor (meeting) conditions for outdoor audio. We then utilized the classifier combination techniques to improve the performance of indoor detectors. The maximum likelihood combination yielded a small gain in performance. We then briefly studied models trained on outdoor conditions; in such a case a large gain in SND performance was observed and the performance of the privacy-sensitive detectors were comparable to the MFCC based detectors.

The chapter concludes with a brief study of speaker diarization on outdoor conditions. While some of the trends that we observed in indoor conditions were validated, the speaker error for even the MFCC features with the SND groundtruth was high. Future studies would ascertain if the results could be improved by parameter tuning or by a scheme compensating for the background noise.

Chapter 7

Summary and Conclusions

Given the emergence of technologies supporting the ubiquitous capture and analysis of real-life audio, in this thesis we have investigated one of the critical issues facing this field, namely that of privacy in the storage of real-life audio. We considered one way to address this issue of privacy; for example, as suggested by (Ellis and Lee, 2004a; Wyatt *et al.*, 2007a) one could store audio features instead of raw audio, such that neither intelligible speech nor linguistic content can be reconstructed. In this context, we investigated four approaches to deriving privacy-sensitive features:

(1) Simple short-time features. One approach to deriving privacy-sensitive features is to extract simple short-time features. Towards this, two sets of features were investigated: (a) energy, zero crossing rate, spectral flatness measure, and kurtosis; (b) Autocorrelation and spectral entropy based features.

(2) Linear prediction residual. Motivated by the source-filter model of speech production, our second approach to privacy was based on linear prediction residual. The source-filter model could be formulated using a mutual-information based framework.

(3) Deep neural network based features. As an extension to the idea of using LP residual, data-driven models assuming the source-filter model could be explored. In this regard, this thesis proposed a deep neural network based method to extract features with low linguistic information — the idea is to train a deep autoassociative neural network with the constraint that only phoneme information is used in the reconstruction. This was achieved by using a deep neural network with a bottleneck architecture, trained using standard backpropagation in a greedy layer-by-layer fashion, to classify phonemes. A reconstruction MLP using the bottleneck output is then trained. The output of this constrained autoassociative neural network is used as a filter to remove the phoneme information.

(4) Obfuscation methods. Obfuscation methods have been used previously for privacy in sensor data research. Here, obfuscation was achieved through either shuffling or averaging feature vectors within non-overlapping blocks of frames.

Using these features, this thesis investigated two aspects of privacy-sensitive features for speech processing: (a) how to quantify and assess the notion of privacy in features; and (b) given features that respect privacy under this assessment, how well would such features perform two tasks, namely speech/nonspeech detection (SND), and speaker diarization. The rest of this chapter summarizes the research carried out, and concludes with some promising future directions.

7.1 Quantitative Privacy Assessment

Qualitative analysis done by two sets of earlier works (Ellis and Lee, 2004a; Wyatt *et al.*, 2007a) indicated that the linguistic message is perhaps the most privacy-sensitive information in the speech signal. A question that this thesis pursued for further investigation was the assessment of the relative privacy of different features. In this context, this thesis proposed phoneme recognition and intelligibility studies, with a higher recognition accuracy and a higher intelligibility being interpreted as lower privacy.

To benchmark privacy, proposed features were compared against MFCC and raw audio. In general, features such as MFCC can be considered to be less privacy-sensitive since they yield an intelligible recovery as well as a state-of-the-art phoneme recognition performance. On the other hand, the proposed approaches yielded lower phoneme recognition accuracies as well as a lower intelligibility, indicating a higher level of privacy.

7.2 Speech/nonspeech Detection

Our study on SND investigated simple feature extraction methods, LP residual features, and obfuscation methods. To evaluate these features, we used the indoor, multiparty conversational meeting data of nearly 450 hours. On this dataset, we evaluated these features and benchmarked them against state-of-the-art spectral shape-based features (MFPLP), on matched and mismatched conditions. To gain further insights, the results were then analyzed for close-talking and far-field microphone scenarios. In addition, this thesis also investigated SND in outdoor conditions. We summarize these results below.

Simple Features

We evaluated the robustness of the two sets of simple features. Explicitly modeling the temporal context is useful in matched and mismatched conditions. Furthermore, combinations of single features yield a bigger gain for the far-field case than the close-talking case. Our studies also show that state-of-the-art performance, comparable to MFPLP features, can be achieved by these simple features for the close-talking scenario. For the far-field, and the mismatched cases, there was still a gap in the SND performance between the simple features and the MFPLP features.

LP Residual Features

Characterizing the excitation source information using LP residual, we showed that exploiting temporal support of up to 51 frames can yield significant gains in the performance. The residual based feature, while performing at only slightly better levels than simple features on close-talking evaluations, performs significantly better when evaluated on far-field data. We also observed that excitation based features are robust not only with respect to distance, but also with respect to mismatched conditions. Fusion strategies combining LP residual with simple features show that state-of-the-art performance can be obtained in both matched and mismatched conditions, on close-talking and far-field microphone scenarios.

Local Temporal Randomization and Averaging

Obfuscation methods using LP residual features caused a small drop in SND performance. However, combinations of randomized or averaged features with simple features yield state-of-the-art SND performance at stricter privacy requirements, defined in terms of phoneme recognition.

Outdoor conditions

The last part of this thesis studied the suitability of using SND models trained on indoor conditions for outdoor conditions. The motivation for this comes from the fact that outdoor and mismatched conditions are more pervasive with real-world data collection. In addition to traditional classifier combination methods such as averaging and inverse entropy to combine indoor detectors, we proposed a maximum likelihood based unsupervised method to combine the detectors. We showed that the proposed combination yields a gain in performance, albeit small. With models trained on outdoor conditions, we confirmed trends that were observed on indoor data.

7.3 Speaker Diarization

Our study on speaker diarization utilized two different approaches to privacy-sensitive audio features for speaker diarization, namely, LP residual based and and deep neural network based. We investigated both sets of features for diarization on the standard meeting (NIST RT06 and RT07) evaluations. Two microphone conditions, namely single and multiple distant microphones (SDM and MDM), were studied. In the last chapter, we briefly investigated the speaker diarization task in outdoor conditions. The results are summarized below.

LP Residual Features on Meetings

We studied two different strategies to represent the LP residual, with the MFCC representation of the residual yielding superior performances for all prediction orders. Additionally, we explored the combination of residual with subband information from 2.5 kHz to 3.5 kHz and spectral slope. Although residual features performed slightly worse than the conventional MFCC features, we observed that residual features are less affected by the change from MDM to SDM scenarios.

Deep Neural Network Features on Meetings

In terms of diarization performance, this approach performed slightly worse than the LP residual based approach. However, these features proved to be more privacy-sensitive then residual features. Future work on this approach will investigate improvements such as training the deep MLP on meeting data.

Outdoor conditions

We set up speaker diarization on the outdoor dataset. Baseline MFCC speaker error was higher than it was in the indoor conditions. Nevertheless, it was reassuring to see that trends similar to those obtained on the indoor data were confirmed. For example, speaker errors made by privacy-sensitive features were slightly higher than those made by MFCC features.

7.4 Recommendations for Future Privacy-Preserving Speech Capture Systems

Based on our studies on SND and speaker diarization, this section offers some suggestions for future implementations of systems that attempt to capture real-life audio while respecting privacy. These suggestions respect the linguistic information in the signal as the most privacy-sensitive information (Ellis and Lee, 2004a; Wyatt *et al.*, 2007a). Specifically, we suggest the following:

1. For applications that exploit the background acoustic information in the signal – for example indexing locations – an approach similar to (Ellis and Lee, 2004a,b) can be utilized, where the wearable device could store randomized versions (over 1 minute) of standard spectral features such as MFCC or PLP.

2. To systems that wish to model speech/nonspeech activity patterns in spontaneous conversations – for example applications that wish to study relationship between speech/nonspeech activities and variables correlated with physical and mental health – we recommend storing the proposed simple features, if the acoustic conditions are clean, such as energy, zero-crossing rate, spectral flatness, and kurtosis. Alternatively, simple features proposed in (Wyatt *et al.*, 2007a) based on autocorrelation and spectral entropy could also be stored. Such features are cheap computationally, and are also privacy-sensitive.

3. On the other hand, systems desirous of processing speech/nonspeech activity patterns in noisy acoustic conditions may need to store MFCC or PLP representations of LP residual. While these features are less privacy-sensitive than the simple features listed above, they are more robust to noisy conditions.

4. If LP residual is stored, and if privacy is a serious issue, we propose an obfuscation approach applied to the residual features with a temporal span of around 13 frames.

5. If speaker activities are to be extracted and then modeled, we suggest storing MFCC or PLP representations of LP residual, in conjunction with measures such as spectral slope and energies derived from high frequency spectral bands. Of course, obfuscation approach can be applied to residual features.

7.5 Future Directions

Some of the promising future directions from this thesis are as follows:

– **Computational analysis:** We remark that there is a tradeoff between privacy, performance, and computational load. While the proposed approaches yield SND and diarization performance comparable to MFPLP and MFCC respectively, these features incur an extra computational load to ensure stricter privacy. Although linear prediction is used in mobile phones (in the form of code excited linear prediction), we would like to assess the computational load of extracting MFCC from LP residual on a wearable device. In this regard, works such as (Lu *et al.*, 2011) have investigated the energy efficiency of implementing speaker identification tasks on a mobile phone. As part of a future line of work, a comparison with such studies in terms of computational complexity and energy efficiency is perhaps necessary.

– **Social acceptability:** Motivated by the qualitative analysis done by (Ellis and Lee, 2004a; Wyatt *et al.*, 2007a), we proposed phoneme recognition and intelligibility studies to investigate the complex issue of assessing privacy in audio. Surveys focused on the social acceptability of privacy in audio are needed not only to ascertain this, but also to determine reasonable norms on the measured phoneme accuracy and the reconstructed intelligibility.

– **Assessment of privacy:** As a first step towards understanding the quantitative assessment of privacy through phoneme recognition and intelligibility studies, we used relatively simple datasets and systems. For example, we used TIMIT in conjunction with the hybrid HMM/MLP system for phoneme recognition. Similarly, we used Semantically Unpredictable Sentences (SUS) from 2 speakers (1 male and 1 female). Studies using more comprehensive systems and realistic datasets could be used for social acceptability studies.

- **Effect of diarization on the analysis of social interactions:** One of the common tools towards the analysis of face-to-face interactions using audio is speaker diarization. Estimates of measures derived from the output of a diarization system, such as speaking time, speaker turns, and overlaps, but also prosodic features can be used to determine social constructs such as dominance, leadership, and roles. In addition, face-to-face affiliation networks can be formed from diarization output to study social roles. Although initial research has been done to determine the relationship between "diarization noise" and dominance (Hung *et al.*, 2011), a systematic investigation of the relation between the diarization noise arising from a privacy-sensitive system and these tasks (dominance, leadership, roles, and the formation of social networks) is needed.
- **Training deep neural networks:** Our training of the deep neural network relied on initializing it using a simple greedy layer-by-layer training approach. While this has been shown to work well, other methods to initialize deep neural networks, such as, autoencoders and restricted Boltzmann machines (RBM) have been shown to yield better results. Given a better trained neural network for phoneme classification (using autoencoders or RBM), we hypothesize that using our strategy to derive features (described in Chapter 5), would yield stricter privacy while yielding better diarization performance. Future work could explore this.

7.6 Conclusions

To summarize, in this thesis we investigated audio features having low linguistic information for two specific tasks, namely speech/nonspeech detection (SND) and speaker diarization. Proposed features were based on four approaches, namely (a) simple features; (b) LP residual features; (c) deep neural network features; and (d) obfuscation approaches. Privacy of the features was quantitatively assessed using automatic speech recognition and intelligibility tests. Proposed features were not only benchmarked against traditional features such as MFCC and PLP on SND and diarization, but also for privacy. Our results show that proposed features preserve privacy better while yielding a performance comparable to the traditional features.

Bibliography

Ajmera, J. and Wooters, C. (2003). A robust speaker clustering algorithm. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, pages 411–416, St. Thomas, U.S. Virgin Islands.

Ajmera, J., McCowan, I., and Bourlard, H. (2004). Robust speaker change detection. *IEEE Signal Processing Letters*, **11**, 649–751.

Anguera, X. (2006). Beamformit, the fast and robust acoustic beamformer. In *`http://www.icsi.berkeley.edu/~xanguera/BeamformIt`*.

Atal, B. S. and Rabiner, L. R. (1976). A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **24**, 201–212.

Basu, S. (2002). *Conversational scene analysis*. Ph.D. thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science.

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2006). Greedy layer-wise training of deep networks. In *Proceedings of Advances in Neural Information Processing Systems*, Vancouver, Canada.

Benoit, C., Grice, M., and Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication*, **18**, 381–392.

Bourlard, H. and Morgan, N. (1994). *Connectionist Speech Recognition- A Hybrid Approach*. Kluwer Academic Publishers, Norwell, Massachusetts, USA.

Bridle, J. (1990). Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In *Proceedings of Advances in Neural Information Processing Systems*. Denver, USA.

Burger, S., MacLaren, V., and Yu, H. (2002). The ISL meeting corpus: The impact of meeting type on speech style. In *Proceedings of International Conference on Spoken Language Processing*, pages 301–304, Denver, USA.

Carletta, J., Ashby, S., Bourban, S., Guillemot, M., Kronenthal, M., Lathoud, G., Lincoln, M., McCowan, I., Hain, T., Kraaij, W., Post, W., Kadlec, J., Wellner, P., Flynn, M., and Reidsma, D. (2006). The AMI meeting corpus: A pre-announcement. In *Proceedings of Workshop on Machine Learning for Multimodal Interaction*, pages 28–39, Edinburgh, Scotland.

Chen., S. and Gopalakrishnan, P. (1998). Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *DARPA speech recognition workshop*, pages 127–132, Virginia, USA.

Choudhury, T. (2004). *Sensing and modeling human networks*. Ph.D. thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science.

Choudhury, T. and Basu, S. (2004). Modeling conversational dynamics as a mixed-memory markov process. In *Proceedings of Advances in Neural Information Processing Systems*, Vancouver, Canada.

Choudhury, T. and Pentland, A. (2003). Sensing and modeling human networks using the sociometer. In *Proceedings of IEEE International Symposium on Wearable Computing*, pages 216–222, White Plains, New York.

Choudhury, T., Clarkson, B., Basu, S., and Pentland, A. (2003). Learning communities: Connectivity and dynamics of interacting agents. In *Proceedings of the International Joint Conference on Neural Networks*, pages 2797–2802, Portland, USA.

Clarkson, B., Sawhney, N., and Pentland, A. (1998). Auditory context awareness via wearable computing. In *Proceedings of Perceptual User Interfaces Workshop*, pages 37–42, San Francisco, USA.

Cover, T. M. and Thomas, J. A. (1991). *Elements of information theory*. Wiley.

Dhananjaya, N. and Yegnanarayana, B. (2007). Speaker change detection in casual conversations using excitation source features. *Speech Communication*, **50**, 153–161.

Dines, J., Vepa, J., and Hain, T. (2006). The segmentation of multi-channel meeting recordings for automatic speech recognition. In *Proceedings of Interspeech*, pages 1213–1216, Pittsburgh, USA.

Donovan, R. (1996). *Trainable speech synthesis*. Ph.D. thesis, Cambridge University, Department of Engineering.

Ellis, D. P. and Lee, K. (2004a). Minimal-impact audio-based personal archives. In *Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, pages 39–47, New York, USA.

Ellis, D. P. W. and Lee, K. (2004b). Features for segmenting and classifying long-duration recordings of personal audio. In *Proceedings of Workshop on Statistical and Perceptual Audio Processing*, Jeju, Korea.

Ellis, D. P. W. and Lee, K. (2006). Accessing minimal impact personal audio archives. *IEEE Multimedia*, **13**, 30–38.

Ernst, M. O. and Buelthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, **8**, 162–169.

Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton, Haag.

Feustel, T. C., Velius, G. A., and Logan, R. J. (1989). Human and machine performance on speaker identity verification. *The Journal of the Acoustical Society of America*, **83**, 169–170.

Fiscus, J. G., Ajot, J., Michel, M., and Garofolo, J. S. (2006). The Rich Transcription 2006 Spring Meeting Recognition Evaluation. In *http://www.itl.nist.gov/iad/mig//publications/storage_paper/RT06SResults-v07.pdf*.

Frankel, J., Wang, D., and King, S. (2008). Growing bottleneck features for tandem ASR. In *Proceedings of Interspeech*, page 1549, Brisbane, Australia.

Furui, S. (1986). Research on individuality features in speech waves and automatic speaker recognition techniques. *Speech Communication*, **5**, 183–197.

Garofolo, J. S., Laprun, C. D., Michel, M., Stanford, V. M., and Tabassi, E. (2004). The NIST meeting room pilot corpus. In *Proceedings of International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Gatica-Perez, D. (2006). Analyzing group interaction in conversations: A review. In *Proceedings of IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 41–46, Heidelberg, Germany.

Gatica-Perez, D. (2009). Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing*, **27**, 1775–1787.

Goldreich, O., Micali, S., and Wigderson, A. (1987). How to play any mental game. In *Proceedings of the nineteenth annual ACM symposium on Theory of computing*, pages 218–229, New York, USA.

Grezl, F. and Fousek, P. (2008). Optimizing bottle-neck features for LVCSR. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pages 4729–4732, Las Vegas, USA.

Grezl, F., Karafiat, M., Kontar, S., and Cernocky, J. (2007). Probabilistic and bottle-neck features for lvcsr of meetings. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pages 757–760, Honolulu, USA.

Hain, T., Burget, L., Dines, J., Garau, G., Karafiat, M., Lincoln, M., Vepa, J., and Wan, V. (2006). The AMI meeting transcription system: progress and performance. In *Proceedings of Workshop on Machine Learning for Multimodal Interaction*, pages 419–431, Bethesda, USA.

Hain, T., Burget, L., Dines, J., Garner, P., Hannani, A. E., Huijbregts, M., Karafiat, M., Lincoln, M., and Wan, V. (2010). The AMIDA 2009 meeting transcription system. In *Proceedings of Interspeech*, pages 358–361, Makuhari, Japan.

Hermansky, H. and Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, **2**, 587–589.

Hermansky, H., Ellis, D., and Sharma, S. (2000). Tandem Connectionist Feature Extraction for Conventional HMM Systems. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pages 1635–1638, Istanbul, Turkey.

Hinton, G. and Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science*, **313**, 504–507.

Hong, J. I. and Landay, J. A. (2004). An architecture for privacy-sensitive ubiquitous computing. In *Proc. of the 2nd international conference on Mobile systems, applications, and services*, pages 177–189, Boston, USA.

Hung, H., Huang, Y., Friedland, G., and Gatica-Perez, D. (2011). Estimating dominance in multiparty meetings using speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, **19**, 847–860.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. (2003). The ICSI meeting corpus. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pages 364–367, Hong Kong.

Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, 226–239.

Krumm, J. (2009). A survey of computational location privacy. *Personal and Ubiquitous Computing*, **13**, 391–399.

Kumaraguru, P. and Cranor, L. F. (2005). Privacy indexes: A survey of westin's studies. Technical Report CMU-ISRI-5-138, Institute for Software Research International (ISRI), Carnegie Mellon University.

Langheinrich, M. (2009a). A survey of RFID privacy approaches. *Personal and Ubiquitous Computing (Special Issue)*, **13**, 413–421.

Langheinrich, M. (2009b). Privacy in ubiquitous computing. *CRC Press*, pages 95–160.

Larochelle, H., Bengio, Y., Louradour, J., and Lamblin, P. (2009). Exploring strategies for training deep neural networks. *Journal of Machine Learning Research*, **10**, 1–40.

Lee, K. and Ellis, D. P. W. (2006). Voice activity detection in personal audio recordings using autocorrelogram compensation. In *Proceedings of Interspeech*, pages 1970–1973, Pittsburgh, USA.

Lee, K. F. and Hon, H. W. (1989). Speaker-independent phone recognition using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **37**, 1641–1648.

Lin, X., Clifton, C., and Zhu, M. (2005). Privacy-preserving clustering with distributed em mixture modeling. *Knowledge and Information Systems*, **8**, 68–81.

Lu, H., Pan, W., Lane, N. D., Choudhury, T., and Campbell, A. T. (2009). SoundSense: scalable sound sensing for people-centric applications on mobile phones. In *Proceedings of the International Conference on Mobile Systems, Applications, and Services*, pages 165–178, Krakow, Poland.

Lu, H., Brush, A. J., Priyantha, B., Karlson, A., and Liu, J. (2011). Energy efficient unobtrusive speaker identification on mobile phones. In *Proc. of the Ninth International Conference on Pervasive Computing (Pervasive 2011)*, pages 188–205, San Francisco, USA.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of IEEE*, **63**, 561–580.

Miller, G. A. (1954). Note on the bias of information estimates. *Information Theory and Psychology*, pages 95–100.

Misra, H., Bourlard, H., and Tyagi, V. (2003). New entropy based combination rules in HMM/ANN multi-stream ASR. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pages 741–744, Hong Kong.

Murty, K. S. R., Yegnanarayana, B., and Guruprasad, S. (2007). Voice activity detection in degraded speech using excitation source information. In *Proceedings of Interspeech*, pages 2941–2944, Antwerp, Belgium.

NIST RT06 (2006). Spring 2006 (RT-06S) Rich Transcription Meeting Recognition Evaluation Plan. http://www.itl.nist.gov/iad/mig//tests/rt/2006-spring/docs/rt06s-meeting-eval-plan-V2.pdf.

NIST RT07 (2007). Spring 2007 (RT-07S) Rich Transcription Meeting Recognition Evaluation Plan. http://www.itl.nist.gov/iad/mig/tests/rt/2007/docs/rt07-meeting-eval-plan-v2.pdf.

OECD (2011). The evolving privacy landscape: 30 years after the OECD privacy guidelines. Technical Report 10.1787/5kgf09z90c31-en, OECD Library.

Olguin-Olguin, D. (2007). *Sociometric badges: wearable technology for measuring human behavior*. Master's thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science.

Olguin-Olguin, D. and Pentland, A. (2010). Sensor-based organisational design and engineering. *International Journal of Organisational Design and Engineering*, **1**, 5–28.

Parthasarathi, S. H. K., Motlicek, P., and Hermansky, H. (2008). Exploiting contextual information for speech/non-speech detection. In *Proceedings of Text, Speech and Dialogue*, pages 451–459, Brno, Czech Republic.

Parthasarathi, S. H. K., Magimai.-Doss, M., Bourlard, H., and Gatica-Perez, D. (2009a). Investigating privacy-sensitive features for speech detection in multiparty conversations. In *Proceedings of Interspeech*, pages 2243–2246, Brighton, United Kingdom.

Parthasarathi, S. H. K., Magimai.-Doss, M., Gatica-Perez, D., and Bourlard, H. (2009b). Speaker change detection with privacy-preserving audio cues. In *Proceedings of International Conference on Multimodal Interfaces and Workshop on Machine Learning for Multimodal Interaction*, pages 343–346, Boston, USA.

Parthasarathi, S. H. K., Magimai.-Doss, M., Bourlard, H., and Gatica-Perez, D. (2010). Evaluating the robustness of privacy-sensitive audio features for speech detection in personal audio log scenarios. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pages 4474–4477, Dallas, USA.

Parthasarathi, S. H. K., Bourlard, H., and Gatica-Perez, D. (2011a). LP Residual Features for Robust, Privacy-Sensitive Speaker Diarization. In *Proceedings of Interspeech*, pages 1045–1048, Florence, Italy.

Parthasarathi, S. H. K., Gatica-Perez, D., Bourlard, H., and Magimai.-Doss, M. (2011b). Privacy-sensitive audio features for speech/nonspeech detection. *To be published in IEEE Transactions on Audio, Speech, and Language Processing*.

Parthasarathi, S. H. K., Bourlard, H., and Gatica-Perez, D. (2011c). Wordless Sounds: Robust Speaker Diarization using Privacy-Preserving Audio Representations. Technical Report Idiap-Internal-RR-29-2011, Idiap Research Institute.

Pathak, M. A. and Raj, B. (2011). Privacy preserving speaker verification using adapted GMMs. In *Proceedings of Interspeech*, pages 2405–2408, Florence, Italy.

Pathak, M. A., Rane, S., and Raj, B. (2010). Multiparty differential privacy via aggregation of locally trained classifiers. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1876–1884, Vancouver, Canada.

Pinto, J., Sivaram, G., Magimai.-Doss, M., Hermansky, H., and Bourlard, H. (2011). Analysis of MLP based hierarchical phoneme posterior probability estimator. *IEEE Transactions on Audio, Speech, and Language Processing*, **2**, 225–241.

Plumpe, M. D., Quatieri, T. F., and Reynolds, D. A. (1999). Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Transactions on Speech and Audio Processing*, **7**, 569–586.

Prasanna, S. R. M., Gupta, C. S., and Yegnanarayana, B. (2006). Extraction of speaker-specific excitation information from linear prediction residual of speech. *Speech Communication*, **48**, 1243–1261.

Prosser, W. L. (1960). Privacy. *California Law Review*, **48**, 383–423.

Sambur, M. (1975). Selection of acoustic features for speaker identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **23**, 176–182.

Smits, R. and Yegnanarayana, B. (1995). Determination of instants of significant excitation in speech using group delay function. *IEEE Transactions on Speech and Audio Processing*, **14**, 325–333.

Solove, D. (2006). A taxonomy of privacy. *University of Pennsylvania Law Review*, **154**, 477–560.

Soong, F. K. and Rosenberg, A. K. (1988). On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **36**, 871–879.

Stupakov, A., Hanusa, E., Bilmes, J., and Fox, D. (2009). A corpus of multi-party conversational speech in noisy environments. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pages 4153–4156, Taipei.

Stupakov, A., Hanusa, E., Vijaywargi, D., Fox, D., and Bilmes, J. (2011). The design and collection of COSINE, a multi-microphone in situ speech corpus recorded in noisy environments. *Computer Speech and Langauge*, **26**, 52–66.

Thevenaz, P. and Hugli, H. (1995). Usefulness of the LPC- residue in text-independent speaker verification. *Speech Communication*, **17**, 145–157.

Vaidya, J., Yu, H., and Jiang, X. (2008). Privacy-preserving SVM classification. *Knowledge and Information Systems*, **14**, 161–178.

Vijayasenan, D. (2010). *An Information Theoretic Approach to Speaker Diarization of Meeting Recordings*. Ph.D. thesis, Swiss Federal Institute of Technology Lausanne (EPFL), Department of Electrical Engineering.

Warren, S. D. and Brandeis, L. D. (1890). The right to privacy. *Harvard Law Review*, **4**, 193–220.

Wester, M. (2010). The EMIME bilingual database. Technical Report EDI-INF-RR-1388, The University of Edinburgh.

Westin, A., the staff of the center for social, and legal research (2003). Bibliography of surveys of the U.S. public, 1970-2003.

Wooters, C. and Huijbregts, M. (2008). The ICSI RT07s speaker diarization system. In *Proceedings of Workshop on Classification of Events, Activities, and Relationships and the Rich Transcription Meeting Recognition*, pages 509–519, Baltimore, USA.

Wrigley, S. N., Brown, G. J., Wan, V., and Renals, S. (2005). Speech and crosstalk detection in multichannel audio. *IEEE Transactions on Speech and Audio Processing*, **13**, 84–91.

Wyatt, D., Choudhury, T., Bilmes, J., and Kautz, H. (2007a). A privacy-sensitive approach to modeling multi-person conversations. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 1769–1775, Hyderabad, India.

Wyatt, D., Choudhury, T., and Bilmes, J. (2007b). Conversation detection and speaker segmentation in privacy-sensitive situated speech data. In *Proceedings of Interspeech*, pages 586–589, Antwerp, Belgium.

Yantorno, R. E. (2000). A study of the spectral autocorrelation peak valley ratio (SAPVR) as a method for identification of usable speech and detection of co-channel speech. Technical report, Speech Processing Lab, Temple University.

Yao, A. C. (1986). How to generate and exchange secrets. In *Proceedings of the 27th Annual Symposium on Foundations of Computer Science*, pages 162–167, Toronto, Canada.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (2000). *The HTK Book Version 3.0*. Cambridge University.

Curriculum Vitae

SHK Parthasarathi

Idiap Research Institute
PO Box 592 CH - 1920
Martigny, Switzerland
Ph: +41 277217731
hari.parthasarathi@idiap.ch

Education

2007 - now	**Ph.D**,
	Swiss Federal Institute of Technology, Lausanne (EPFL) and Idiap Research Institute,
	Advisors: Dr. Daniel Gatica-Perez and Prof. Herve Bourlard,
	Thesis : Privacy-Sensitive Audio Features for Conversational Speech Processing.
2005 - 2007	**M.S (Computer Science - CGPA: 9.8/10)**,
	Indian Institute of Technology (IIT Madras),
	Advisor: Prof. Hema A Murthy,
	Thesis: Voice Activity Detection Using Group Delay Functions.

Professional Experience

2007–2011 Research Assistant at Idiap Research Institute:

A key impediment to the ubiquitous capture of audio using mobile phones is the issue of privacy. Towards this, we study audio features that can achieve state-of-the-art performance in *speech/nonspeech detection* and *speaker diarization* while minimizing the amount of linguistic information. A comprehensive analysis of various features for these tasks is performed in indoor (predominantly) and outdoor conditions. To objectively evaluate the notion of privacy, we perform *human and automatic speech recognition tests*.

2001–2004 Senior Software Engg at Honeywell Bangalore:

I was with 3 departments of Honeywell: home and building, research and technology, and avionics. I was involved with, (a) Design and implementation of a real-time data retrieval system for the Airbus A380; (b) Investigation and prototyping of a text classification system using aircraft log data; (c) Design and implementation of a compiler and a virtual machine for a rule based language to interface with Honeywell devices; and (d) Design and implementation of an inmemory DBMS engine.

List of Publications

Journals Publications

1. **SHK Parthasarathi**, D Gatica-Perez, H Bourlard, and M Magimai.-Doss, "Privacy-Sensitive Audio Features for Speech/Nonspeech Detection", To be published, *IEEE Transactions on Audio, Speech, and Language Processing 2011.*
2. **SHK Parthasarathi**, H Bourlard, and D Gatica-Perez, "Wordless Sounds: Robust Speaker Diarization using Privacy-Preserving Audio Representations", under revision, *IEEE Transactions on Audio, Speech, and Language Processing.*
3. **SHK Parthasarathi**, R Padmanabhan, and HA Murthy, "Robustness of Group Delay Representations for Speech Signals", *IJST, Springer-Verlag 2011.*

Conference Publications

1. **SHK Parthasarathi**, H Bourlard, and D Gatica-Perez, "LP Residual Features for Robust, Privacy-Sensitive Speaker Diarization", *Proc. of Interspeech, 2011.*

2. **SHK Parthasarathi**, M Magimai.-Doss, H Bourlard, and D Gatica-Perez, "Evaluating the Robustness of Privacy-Sensitive Audio Features for Speech Detection in Personal Audio Log Scenarios", *Proc. of ICASSP, 2010.*

3. **SHK Parthasarathi**, M Magimai.-Doss, D Gatica-Perez, and Herve Bourlard, "Speaker Change Detection with Privacy-Preserving Audio Cues", *Proc. of ICMI-MLMI, 2009.*

4. **SHK Parthasarathi**, M Magimai.-Doss, H Bourlard, and D Gatica-Perez, "Investigating privacy-sensitive features for speech detection in multiparty conversations", *Proc. of Interspeech, 2009.*

5. R Padmanabhan and **SHK Parthasarathi** and HA Murthy, "Robustness of phase based features for speaker recognition", *Proc. of Interspeech, 2009.*

6. **SHK Parthasarathi**, P Motlicek, and H Hermansky, "Exploiting contextual information for speech/non-speech detection", *Proc. of TSD 2008, LNCS/LNAI series, Springer-Verlag.*

7. R Padmanabhan and **SHK Parthasarathi** and HA Murthy, "A Pattern Recognition Approach to VAD using modified group delay", *Proc. of National Conference on Communications (NCC), 2008.*

8. YR Venugopalakrishna and **SHK Parthasarathi**, S Thomas, K Bommepally, K Jayanthi, H Raghavan, S Murarka, and SA Murthy, "Design and Development of a Text-To-Speech Synthesizer for Indian Languages", *Proc. of National Conference on Communications (NCC), 2008.*

9. **SHK Parthasarathi**, R Padmanabhan, and HA Murthy, "Voice Activity Detection using Group Delay Processing on Buffered Short-term Energy", *Proc. of National Conference on Communications (NCC), 2007.*

10. **SHK Parthasarathi**, R Padmanabhan, and HA Murthy, "Robust Voice Activity Detection using Group Delay Functions", *Proc. of IEEE International Conference on Industrial Technology (ICIT), 2006.*