Wordless Sounds: Robust Speaker Diarization using Privacy-Preserving Audio Representations

Sree Hari Krishnan Parthasarathi *Student Member, IEEE*, Hervé Bourlard *Fellow, IEEE*
and Daniel Gatica-Perez *Member, IEEE*

***Abstract*—This paper investigates robust privacy-sensitive audio features for speaker diarization in multiparty conversations: ie., a set of audio features having low linguistic information for speaker diarization in a single and multiple distant microphone scenarios. We systematically investigate 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 Mel Frequency Cepstral Coefficients (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.**

***Index Terms*—Privacy sensitive audio features, speaker diarization, LP residual, deep neural networks, listening tests.**

I. Introduction

OUR work takes place in the context of analyzing social interactions using multimodal sensors with an emphasis on audio [1]. Towards this we wish to capture spontaneous conversations using portable audio recorders. Analysis of conversations can then proceed by modeling the speech/speaker activities produced by a speaker diarization system. Traditionally, diarization is a batch process without any prior knowledge of the speakers [2].

However, recording and storing raw audio for this purpose would breach the privacy of people whose consent has not been explicitly obtained [3]. Some studies have suggested that the linguistic message is the most privacy-sensitive information ([3], [4]). To respect this notion of privacy, features could be stored from which neither an intelligible speech nor the lexical content can be reconstructed. We take this approach to extract privacy-sensitive features and then to apply diarization.

While the output of a diarization system may appear to be restrictive, there are a growing number of applications that model the speech/speaker activities (derived from diarization) for studying productivity and personal health. For example, [5] presents several case studies investigating organizational productivity using such measures, implemented on wearable devices. In the medical community, [6] uses a portable device to record features from which, among others, speech activity was extracted to study physical and mental health.

Apart from privacy, another constraint in recording audio using wearable devices is that the features for diarization be robust to single distant microphones (SDM). Traditional meeting room diarization, in contrast, uses multiple distant microphones (MDM) [7]. This paper focuses on the former, exploring the tradeoff between diarization performance and audio privacy.

State-of-the-art diarization systems [7], use features derived from the spectral shape such as Mel Frequency Cepstral Coefficients (MFCC). While these features are relatively robust to SDM, Milner et al. [8] show that highly intelligible speech can be reconstructed from MFCC. Previous approaches to privacy-sensitive features have focused on either reinterpreting simple, frame-level heuristics for estimating speech activity in conversations [4], [9], or computing long-term averages of standard features for indexing audio logs [3]. However these methods were not proposed for diarization, a choice that is further supported by results in our preliminary experiments.

In this paper, drawing motivation from the source-filter model of speech production, we investigate linear prediction (LP) residual for diarization. Besides prediction order, two different representations of LP residual are compared, namely, real-cepstrum and MFCC, with the latter yielding better performance. We explore the combination of residual with subband information (2.5 kHz to 3.5 kHz) and spectral slope. To enforce stricter privacy, we study obfuscation methods such as local temporal randomization [10] of residual features.

In addition to LP residual, we propose a supervised residual, obtained using a deep neural network (DNN) with a bottleneck architecture. We benchmark both features against MFCC using the diarization system presented in [11] on the NIST RT07 dataset [12]. Results show that the proposed features yield performances close to MFCC in SDM condition.

The notion of privacy in audio remains something that is difficult to evaluate. Exploiting studies suggesting the linguistic information as the main privacy concern ([3], [4]), this paper presents human speech recognition (HSR) and phoneme recognition to assess privacy, with higher accuracy

S.H.K Parthasarathi is with Idiap Research Institute, P.O. Box 592, CH-1920 Martigny, Switzerland and Ecole Polytechnique Fédérale, Lausanne (EPFL), Switzerland e-mail: hari.parthasarathi@idiap.ch

H. Bourlard is with Idiap Research Institute, P.O. Box 592, CH-1920 Martigny, Switzerland and Ecole Polytechnique Fédérale, Lausanne (EPFL), Switzerland e-mail: herve.bourlard@idiap.ch

D. Gatica-Perez is with Idiap Research Institute, P.O. Box 592, CH-1920 Martigny, Switzerland and Ecole Polytechnique Fédérale, Lausanne (EPFL), Switzerland e-mail: gatica@idiap.ch

being interpreted as lower privacy. We show that the proposed approaches are more privacy-sensitive than MFCC.

The contributions of this paper are: (a) a systematic investigation of LP residual features for diarization in SDM and MDM conditions; (b) a DNN architecture for extracting features; and (c) evaluation of privacy in audio using HSR and phoneme recognition. The findings of this paper are that the proposed features yield a diarization performance close to MFCC on SDM, while yielding much stricter privacy.

The rest of the paper is organized as follows. Section II reviews the literature on LP residual and DNN. An overview of our methodology is summarized in Section III. A description of the proposed features is given in Section IV, while Section V discusses the diarization setup. Parameters selection experiments associated with the proposed features is described in Section VI. Results are presented in VII and Section VIII. Finally, conclusions are drawn in Section IX.

II. Related work

So far, we have discussed the relevant work on privacy-sensitive features. In this section, we briefly survey related work in LP residual and deep neural networks.

A. Linear prediction residual

It is generally known that up to three formants are required to synthesize intelligible speech or to reconstruct the linguistic information [13]. Motivated by the source-filter model, our approach to preserving privacy is based on adaptively filtering these spectral peaks.

LP analysis of speech [14] assumes the source-filter model and it estimates three components, namely an all-pole model, a residual and 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 residual, obtained by inverse filtering the signal with the all-pole model, can be considered to be privacy-preserving. Depending on LP order, residual contains information mostly about the excitation source of speakers [15]. It has also been shown that humans can recognize speakers by listening to the residual signal [16].

Previous works have exploited the speaker information in LP residual. For example, ([15], [17]) use residual for speaker recognition. In an earlier work [18], interpreting the LP order as a tradeoff between privacy and speaker information, we explored LP residual as a feature for speaker change detection. To our knowledge, this is the first work exploiting residual for diarization in SDM and MDM scenarios.

In sensor data research, methods of obfuscating data to preserve privacy, such as randomization, are well established [10]. Persuaded by obfuscation methods, we conjecture that while temporal dynamics of the speech signal is important for its intelligibility, it could be less so for speaker recognition tasks; local temporal randomization (within 250 ms) of residual features is explored.

B. Deep neural networks (DNN)

Our interest in DNN stem from its suitability for representing phonemes. This section reviews DNN, while a later section (IV-B), describes the proposed feature extraction using DNN.

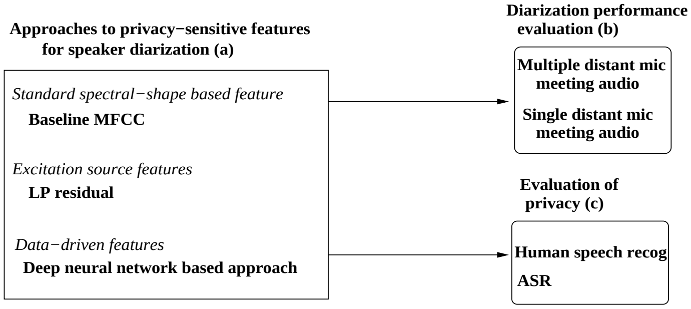
Approaches to privacy−sensitive features for speaker diarization (a)
Standard spectral−shape based feature
Baseline MFCC
Excitation source features
LP residual
Data−driven features
Deep neural network based approach
Diarization performance evaluation (b)
Multiple distant mic meeting audio
Single distant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

Fig. 1. Block diagram of our approach. A detailed discussion of the figure is provided in Section III.

Feedforward neural networks with a 3-layer architecture, also called multilayer perceptrons (MLP), have been used for feature extraction in automatic speech recognition (ASR) for several years [19], [20]. Recently, DNN i.e., the number of layers more than three (alternatively, number of hidden layers more than one), are receiving attention due to their ability to represent knowledge in a principled fashion ([21], [22]). The motivation comes from the complexity theory of circuits [23].

Of interest to this work are DNN with bottleneck architectures, which are recently explored for ASR in the quest towards obtaining better phoneme representations. For example, [22] shows that the output from the bottleneck layer of a trained 5-layer MLP yields lower word error rates in comparison to the traditional probabilistic features.

A key issue with DNNs is the difficulty in training its weights. A gradient-based optimization starting from random initialization has been reported to get trapped in local optima [23]. This was also observed by us while training networks with more than three layers for phoneme recognition on TIMIT, to the extent that deeper networks perform worse. Two common strategies to address this difficulty are, greedy layer-by-layer training [24], [25], and an autoencoder training [21].

Since the privacy constraints imply that the derived features cannot capture phonemes, we deploy a reconstructed spectrum from the bottleneck layer could be deployed as an inverse filter and hypothesize that it yields a privacy preserving representation. Section IV-B describes this in detail.

III. Overview of our methodology

This section, composed of three stages, summarizes our overall methodology in Figure 1.

(a): We begin with a detailed description of the features extracted from LP residual and DNN in Sections IV-A and IV-B. To gain further insight, Section IV-C provides a more formal analysis using mutual information.

(b): Benchmarking privacy-sensitive features entails a comparison of diarization performance as well as linguistic privacy. Details of the diarization system, features, datasets, and the baseline performance are presented in Section V. Parameter selection for the proposed features on the development (SDM and MDM) data is discussed in Section V. Diarization results are presented in Section VII. We discuss the MDM scenario mainly as a reference to the existing literature.

(c): Experimental protocol and the results for HSR and phoneme recognition are provided in Section VIII.

IV. Privacy-sensitive features

We present the details in deriving the proposed features and follow that by an analysis based on mutual information.

A. LP residual features

We discuss features derived from LP residual, subband information, and spectral slope.

(a) LP residual: LP residual is extracted every 10 ms, using a hamming window of size 30 ms. Two representations of the residual studied are: real-cepstrum ([17]) and MFCC with 19 coefficients each. These representations have been fixed at 19 dimensions to have the same dimensions as the baseline MFCC features. The MFCC representation is computed using HTK [26]. Feature selection experiments analyzing both representations are presented in Section VI. Effect of the prediction order, representing a tradeoff between privacy and performance, is studied by varying it from 2 to 20.

(b) Subband information: Previous studies have shown that the spectral subband, 2500 Hz to 3500 Hz, carries speaker specific information [27]. In an earlier study [18], we exploited this for speaker change detection (SCD) by representing the subband using three MFCC. An MFCC representation decorrelates the filterbank energies and makes it suitable for a Gaussian Mixture Model (GMM) with diagonal covariance matrices. To compute subband MFCC, we employed HCopy [26]: it bandlimits the signal between 2500 Hz to 3500 Hz, and distributes the four filterbank channels equally on the mel scale such that the lower cutoff of the first filter is at 2500 Hz and the upper cutoff of the fourth filter is at 3500 Hz. Three cepstral coefficients are then calculated from the four values using Discrete Cosine Transform (DCT).

(c) Spectral shape: Generally, speakers differ in the distribution of spectral energies [28]. For instance, male and female speakers exhibit different spectral energy distribution. Spectral slope (SS) is a way to characterize this, with the spectrum of female speakers tending to show a steeper slope than male speakers. In [18] we showed that the first cepstral coefficient (c_1) obtained from LP analysis can enhance SCD when combined with the residual features.

(d) Obfuscation (local temporal randomization): Features within non overlapping blocks of sizes ($N = 1, 5, 9, 13$) are shuffled using a uniform pseudo-random number generator. Such a randomization could result in two successive frames being separated by $2 \cdot (N - 1)$ frames. The choice of the upper limit for N being 13 frames was guided by results from [29], which indicate that information in the speech signal up to 230 ms can be exploited for phoneme recognition.

B. DNN 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.

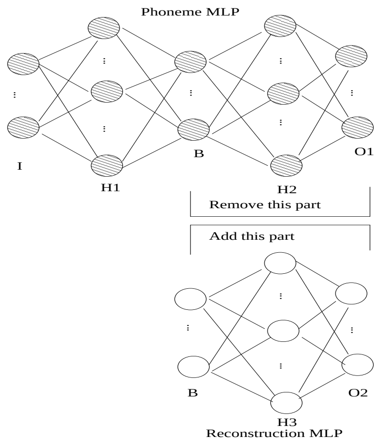

Fig. 2. 5-layer deep neural network with bottleneck architecture. (a) 5-layer phoneme MLP is trained with phoneme targets using cross entropy criterion (b) Keeping weights for the first 2 layers fixed, and removing last 2 layers, a reconstruction MLP is trained for the last two layers with squared error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

1) 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 [22], [24], 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 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 [30] 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 [22], and we varied the number of units from 20 to 40 [30].

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 [25]. Cascaded MLPs with 3, 4, and 5 layers are trained using standard back propagation algorithm by minimizing the cross entropy error criterion ([22], [24]). 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 [29].

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

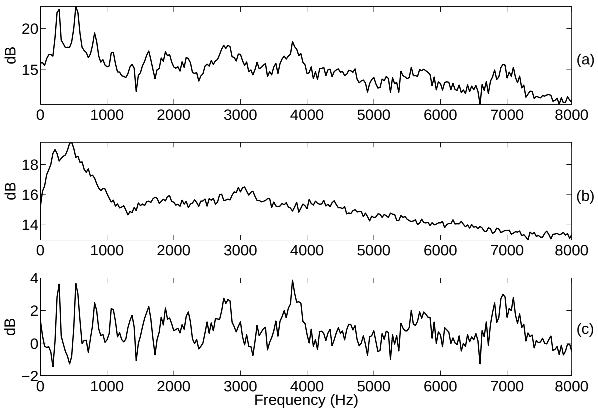

Fig. 3. Example steps in neural network filtering for an input frame that is /iy/: (a) Input to phoneme MLP (logarithm of DFT square magnitude vector) (b) Output from reconstruction MLP (logarithm of DFT square magnitude vector) (c) Filtered spectrum.

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 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.

2) 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 (Section VI) 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.

3) An example: Figure 3 illustrates the example steps in neural network filtering for an input frame that is /iy/ phoneme. Figure 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 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 3(c) shows the filtered spectrum. We observe that the spectral shape (mainly the first formant) is filtered.

C. Mutual information based analysis

We now present a discussion on privacy using mutual information (MI). Privacy in audio could be interpreted as a function that maximizes the MI with speakers while minimizing the MI with phonemes. This is followed by an analysis on the TIMIT test data (1344 utterances from 168 speakers).

1) 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 $I(g(X);S) - I(g(X);Q)$.

$$g^* = \arg\max_g I(g(X);S) - I(g(X);Q) \qquad (1)$$

This equation is, in general, difficult to solve without additional assumptions. Assuming that Q and S are independent[1], the maximum of Eq (1) is reached for:

$$g^*(X) = \tilde{S} \qquad (2)$$

where $\tilde{S}$ is a transformation of X that has maximum MI with S. A further assumption of a source-filter model of speech production simplifies this to:

$$g^*(X) = \tilde{S} = X - \tilde{X} \qquad (3)$$

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 Eq 3.

Deep neural network filter: An alternative is to train a data-driven filter that yields $\tilde{X}$, given X as input. 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,

$$Z = \psi(X;\theta_1,\theta_2,\mathcal{D}) \qquad (4)$$

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-

[1] It might be that speakers can have biases towards choices of words and therefore towards phoneme

entropy training criterion can be written as:

$$\begin{aligned}\mathcal{J}(\theta_1,\theta_2) &= -E_X[\sum_k P(q_k|x)\log\tilde{P}(q_k|x)]\\ &= -\int_X p(x)\sum_k P(q_k|x)\log\tilde{P}(q_k|x)dx\\ &= -\int_X \sum_k P(q_k,x)\log\frac{\tilde{P}(q_k|x)\tilde{P}(x)\tilde{P}(q_k)}{\tilde{P}(x)\tilde{P}(q_k)}dx\\ &= -\int_X \sum_k P(q_k,x)[\log\frac{\tilde{P}(q_k,x)}{\tilde{P}(x)\tilde{P}(q_k)}+\log\tilde{P}(q_k)]dx\\ &= I(Q;X)-\sum_k P(q_k)\log\tilde{P}(q_k) \qquad (5)\end{aligned}$$

It can be seen from the above equation that minimum cross-entropy training is equivalent to maximum mutual information training [31]. Let B denote the random variable obtained at output from the bottleneck layer before the nonlinearity. Then,

$$B=\phi(X;\theta_1,\mathcal{D}) \qquad (6)$$

where θ_1 is the set of parameters of the MLP up to the bottleneck layer. Furthermore, from data-processing inequality [32],

$$I(X;Q)\geq I(B;Q)\geq I(Z;Q) \qquad (7)$$

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. Together with the fact that the dimension of the bottleneck (B) is much smaller than the dimension of input (X), means that bottleneck (B) serves as a compression of input (X) retaining information that has maximum MI with the phonemes (Q).

It is, therefore, reasonable to assume that other information such as speakers (S) is lost at bottleneck. We now consider the reconstruction MLP, which is trained with bottleneck (B) as input, and X as the training target, by minimizing the squared error. 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 MI 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 $\tilde{X}$ in Eq (3), we obtain $\tilde{S}$.

2) MI analysis: In practice, we can introduce a variable (λ) in Eq (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. We take the latter approach and Figure 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 [33].

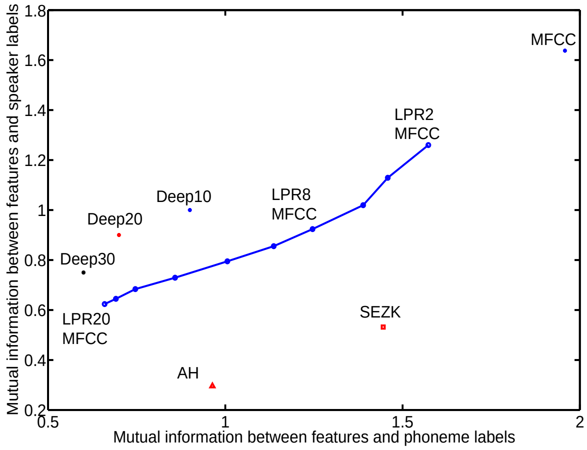

Fig. 4. Plot showing MI between the features and phonemes versus MI between the features and speakers. LPRx denotes residual features with LP order x. *SEZK* and *AH* denote the features from [9] and [4] respectively. Deepxy refer to DNN features with bottleneck sizes corresponding to xy.

Figure 4 plots baseline MFCC, residual, and DNN 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 speech [13].

For the DNN 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 DNN features (with 20 bottleneck units) yield much lower MI with phoneme labels, while yielding similar MI with speaker labels.

Features from [9] and [4] 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.

V. Diarization Setup

This section discusses the diarization system, features, datasets and the performance measure.

A. Diarization system

The diarization system is based on ergodic Hidden Markov Model (HMM) as described in [11], where each 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 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 [34]. At each step, all possible cluster merges are compared using a modified version of the BIC criterion [11].

This HMM/GMM based diarization system uses the baseline 19 dimensional MFCC features, which are extracted every 10 ms, with a hamming window of size 30 ms using HTK [26]. Delta and acceleration features are not used.

B. Privacy-sensitive features

The proposed privacy-sensitive features are compared against the baseline 19 dimensional MFCC using the system discussed in Section V-A. To summarize Section IV, 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 [11].

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

C. Datasets

Experiments were performed on NIST RT06 and RT07 evaluation data for Meeting Recognition Diarization task [12], [35]. 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 [36]. 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 [37]. Since our interest in this paper is in evaluating the privacy-sensitive features for speaker segmentation and clustering, the same speech/nonspeech segmentation is used across all experiments.

D. 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 paper, we include the overlapped speech in the evaluation.

TABLE I
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 (%) MDM	Spkr err (%) SDM
RT06	6.5	0.1	6.6	**17.1**	**20.8**

Table I 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 in terms of the speaker error for both MDM and SDM scenarios. As expected, MFCC performs better on the MDM data. On RT06 we observe a performance gain of 3.7% on MDM over SDM.

VI. Parameter selection on RTeval06

Recall that we use RTeval06 as the development dataset. In Section IV-C, 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 diarization system on RTeval06.

A. LP residual features

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

1) Representations of LP residual: We study the 2 different representations of LP residual using the baseline diarization system described in Section V-A. Figure 5 shows the comparison between the 2 representations on the RT06 MDM 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.

2) Prediction order: The effect of LP order on MFCC representation of residual on both MDM and SDM data is presented in Figure 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.

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

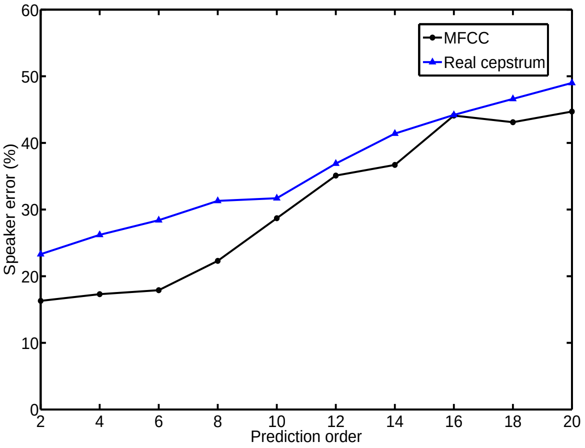

Fig. 5. Comparison between MFCC and real-cepstrum representations of the LP residual on RT06 MDM evaluation data.

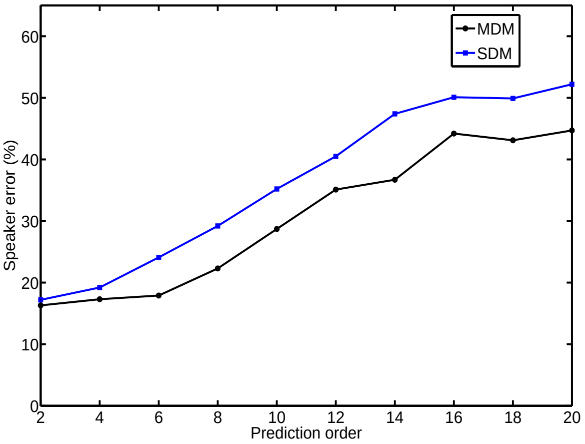

Fig. 6. Using MFCC representation of LP residual, prediction order vs speaker error is illustrated on MDM and SDM conditions of the development dataset (RT06).

case. This could be due to the loss of the first formant, which carries more linguistic information [13]. 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 to 5 formants. Since higher order formants carry more speaker information [38], we note that increasing prediction order beyond 8 results in greater speaker errors.

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

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, an LP order of 8 seems appropriate for the diarization task, since the first two formants are important for synthesizing an intelligible speech signal [13]. At this prediction order, residual yields a performance of 22.3% on the MDM data while yielding 29.2% on the SDM data.

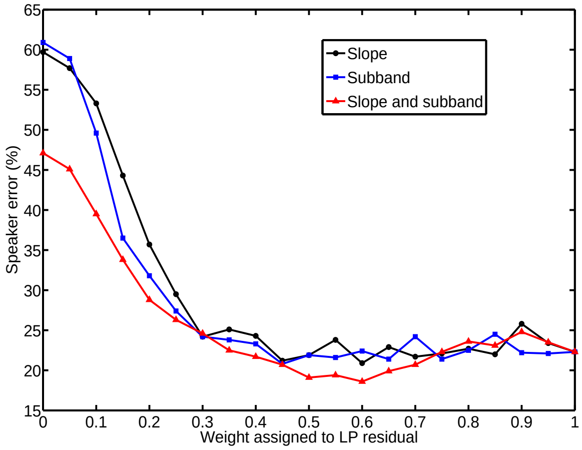

Fig. 7. Combination of LP residual (MFCC representation) with slope and subband. X-axis denotes the weight assigned to LP residual.

3) 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 7. X-axis denotes the weight assigned to LP residual, while 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, while a weight of 0 denotes that these features are used without LP residual.

It can be observed from the plot that for either slope or subband energies, combining residual with weights less than 0.45 yields a lower performance than that is 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 an error of 20.7% at a weight of 0.45, while the best combination with subband energy yields an error of 20.9% at a weight of 0.6.

We note that combining both slope and subband energies yields a consistent gain over combining with either of those features. Furthermore, combining both features with residual yields improvement over residual by itself, for weights between 0.45 to 0.8. The best performance of this combined system is 18.6% 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.

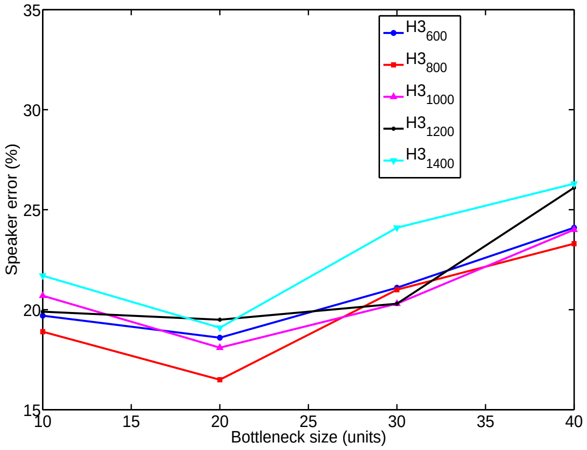

Fig. 8. Performance of DNN features on the development data. Bottleneck size (B - in terms of number of units) versus speaker error rates (%) for 5 different reconstruction layer sizes (H3). The input features are squared magnitude vectors.

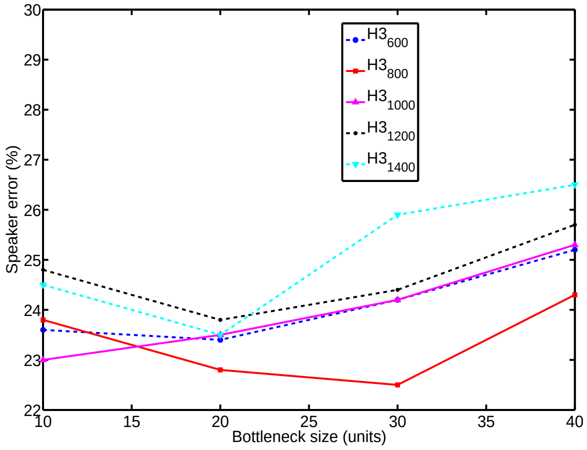

Fig. 9. With input features as MFCC, performance of the DNN. Bottleneck size (B - in terms of number of units) versus speaker error rates (%) for 5 different reconstruction layer sizes (H3).

B. DNN features

We now analyze the parameters of the DNN approach, 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 log squared magnitude vectors were obtained on the MDM development data (RT06 eval). MFCC representation was obtained from the log squared magnitude vectors and the ICSI diarization system was used to analyze the features.

Figures 8 and 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.

1) Log squared magnitude input: For the experiments in Figure 8, the input to the phoneme MLPs was 257 dimensional log squared magnitude vectors. The output of the reconstruction MLP was 257 dimensional log 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 that 1000 nodes to be a reasonable choice for the first and third layers of the phoneme MLP.

From Figure 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 [30]. 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 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 DNN 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.

2) MFCC input: We now examine Figure 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.

3) Filtering domain: We performed studies on MFCC being the output of the reconstruction MLP. Diarization experiments showed that the speaker error was high. Since the objective of the paper was not to optimize all the parameters of the proposed DNN features, 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.

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

VII. Diarization Results on RTeval07

Recall that we use the HMM/GMM based diarization system [11], and that we utilize the MDM and SDM conditions in RTeval07 for evaluation. This diarization system is used to evaluate the the proposed privacy-sensitive features against the baseline MFCC features.

A. Baseline MFCC

This section begins with the results obtained using the baseline MFCC features, which are tabulated in Table II. 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 II
Performance of baseline MFCC features on RT07 MDM and SDM data: The first 3 columns list the performance of the speech/nonspeech detection while the next 2 columns report the speaker errors.

Features	Miss	FA	sp/nsp	Spkr err (%) MDM	Spkr err (%) SDM
MFCC (baseline)	3.7	0.0	3.7	**6.4**	**11.2**

B. Comparison with MFCC on RT07 MDM

Table III 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 DNN features summarized in Section VI-B4.

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 [2].

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 III 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 in Section VII-D.

C. 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 III.

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 [38] – that are removed by LP analysis as well as by DNN. These conclusions are supported by our results for speaker change detection in [18], 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).

TABLE III
RT07 evaluation data: Performance of 8^{th} order LP residual and DNN 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

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 [39]. 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 III, 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 3 shows that the DNN approach captures pitch information.

D. Meetingwise comparison

Table IV 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 IV
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 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 entire dataset.

On MDM, not only does MFCC perform better than residual and DNN features on the whole data, it performs better on

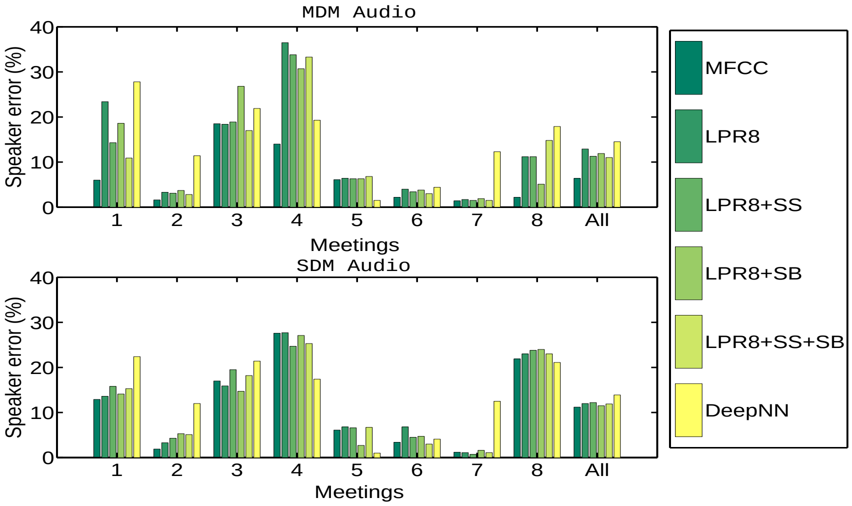

Fig. 10. Meetingwise analysis of the 9 meetings in the RT07 evaluation dataset. The upper plot shows the comparison on the MDM audio while the lower plot shows the comparison the SDM audio. The meeting numbers correspond to the first column in Table IV.

most meetings. This supports our analysis in Section VII-C. 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.

E. Obfuscation method

In Section IV-A, we mentioned another strategy that can be gainfully employed for improving privacy of audio features. We now present speaker error rates of MFCC and LPR8 that are randomized with block sizes ($N = 5, 9, 13$) on the evaluation dataset in Table V. 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.

TABLE V
Effect of randomization on MFCC and LPR8 on the RT07 MDM dataset. Randx is used to denote randomization with block size of x frames. Baseline, non randomized performances are given as a reference in the first row.

Feature	LPR8 (%) Spkr err	MFCC (%) Spkr err
Baseline	12.9	6.4
Rand5	13.4	6.7
Rand9	13.8	7.1
Rand13	13.7	6.8

VIII. Analysis of privacy

So far, we have investigated LP residual and DNN features for speaker diarization. We now proceed to analyze privacy.

To our knowledge, quantitative analysis of privacy in audio has not been studied before. Studies such as [3], [4] indicate that the main privacy concerns are the reconstructibility of an intelligible speech signal and of the linguistic information. In this paper, we explore two ways to analyze this notion of privacy: human speech recognition (HSR) rates of speech synthesized from the features and automatic speech recognition (ASR) rates using the features. ASR accuracies are generally reported in the literature using phoneme recognition or word recognition. Since the latter is more complex for assessing privacy due to the differences in vocabulary sizes, dictionaries, and language models, we prefer phoneme recognition studies.

A. Analysis using human speech recognition

In the field of HSR, one aspect of an intelligibility test is whether the vocabulary is open or closed. Another aspect is whether one tests on individual units such as nonsense syllables or on fully-formed sentences. Furthermore, fully-formed sentences could be meaningful such as conversations and news or semantically unpredictable sentences (SUS) [40].

In this study, we selected a dataset that was open vocabulary as well as being SUS. This is done so that the test evaluates the acoustic aspect of intelligibility instead of the cognitive aspect of prediction. SUS are usually constructed from simple grammatical templates.

1) HSR setup: We used the 20 SUS from EMIME bilingual database [41], with a vocabulary size of 88 words. The list of sentences is given in Table VI. 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.

TABLE VI
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.

We generated the following features from this audio: (a) baseline MFCC; (b) MFCC representation of 8^{th} order LP residual; and (c) MFCC representation of DNN features. Reconstruction[2] yields 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 (or 4^{th} *system*) to estimate the upper bound in performance.

In the tradeoff between obtaining reasonable estimates of intelligibility versus repeating each sentence, we divided the 80 utterances (20 sentences $\times$ 4 systems) into 2 groups of 40 each. Each group was obtained by a Latin square design to maximize the coverage of the systems and the sentences. In order that listeners do not get used to a predetermined sequence of audio, the sequences were randomized. Each listener was assigned to one of the two groups and she listened to 40 utterances (10 utterance from each system).

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 5 listening tests, they typed "Not intelligible". Out of the 27 listeners, one was a native english listener.

2) HSR experiments: Before scoring, we preprocessed the listeners' typed-in responses to ensure that typing errors are not counted as a loss in intelligibility. The score, computed using the HResults tool [26], is the ratio of the number of correct words to the total number of words.

The results of scoring the features are listed in Table VII. In addition, we also obtained an ordering of listeners according to the percentage of words correctly recognized. In Table VII, 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. 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 DNN features.

TABLE VII
HSR performance 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 two formants, which carry more linguistic information [13]. There is a further loss in information by representing LP residual using MFCC. DNN features yield the lowest intelligibility, 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.

B. Analysis using automatic phoneme recognition

Another approach to assessing linguistic privacy is to study phoneme recognition accuracies for privacy-sensitive and MFCC features. 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 [29].

1) 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. For all the features studied (baseline MFCC, LP residual with MFCC representation, DNN features with MFCC representation), the input to the MLP was 13-dimensional MFCC with delta and acceleration coefficients. The MLP

[2] We obtained a noise-excited reconstruction from MFCC using the RASTA-MAT library: http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

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 [19]. 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 same, and is derived from the output of the MLP.

2) Phoneme recognition experiments: Figure 11 plots the recognition accuracies for increasing LP orders. 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 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 [9], namely, spectral flatness, energy, zero-crossing rate, and kurtosis (*SEZK*) and the features proposed in [4], 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 DNN features yielded an accuracy of 48.7%, which is much lower than that of 8^{th} order LP residual's.

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.

IX. Conclusion

In this paper we presented two different approaches to privacy-sensitive audio features for robust speaker diarization, namely, LP residual based and and DNN 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 was interpreted as the linguistic message, and methods to assess them in terms of phoneme recognition and intelligibility tests were studies. We now summarize our key conclusions.

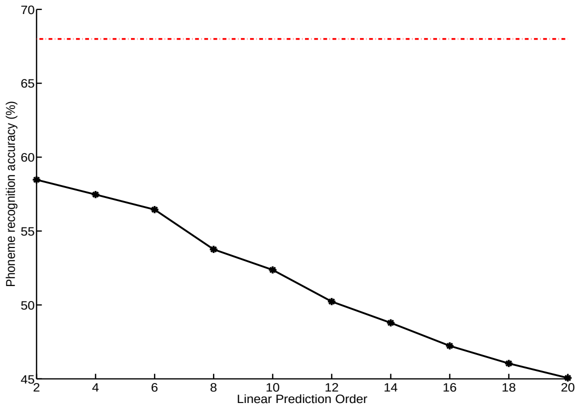

Fig. 11. *Phoneme recognition accuracy for the residual based features various LP orders on TIMIT. The x-axis shows the LP order while the y-axis shows the phoneme accuracy in (%).*

1) LP residual features: 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. Furthermore, residual features proved to be more privacy-sensitive than MFCC features in terms of lower intelligibility and phoneme recognition accuracy.

2) DNN features: We utilized a greedy, layer-by-layer trained DNN for representing the phoneme information in the short-term spectrum of the signal. A second MLP was utilized to reconstruct the spectrum, which was used as a 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 DNN on meeting data.

3) Putting privacy and diarization together: 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 on SDM, the performance of the DNN features were about 2% lower than MFCC. 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 [42].

4) Future Work: Nonverbal cues in audio have been explored in developing computational models of face-to-face human behavior. However, with a few exceptions [5], [6], most work done in this domain are from meeting room audio. Our future work will utilize the privacy-sensitive audio features in this paper to capture real-world audio. Patterns of speech/nonspeech detection and diarization can then be used to analyze social interactions.

Finally, in this paper, we have proposed intelligibility and phoneme recognition as means to investigate the complex issue of assessing privacy in audio. Complementary social acceptability studies are needed to determine reasonable norms on measured phoneme accuracy.

Acknowledgment

This work was supported by the Swiss National Science Foundation through the project Interactive Cognitive Systems (ICS).

References

[1] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image and Vision Computing*, vol. 27, pp. 1775–1787, 2009.

[2] D. Vijayasenan, "An Information Theoretic Approach to Speaker Diarization of Meeting Recordings," Ph.D. dissertation, Swiss Federal Institute of Technology Lausanne (EPFL), Department of Electrical Engineering, 2010.

[3] D. P. W. Ellis and K. Lee, "Accessing minimal impact personal audio archives," *IEEE Multimedia*, vol. 13, pp. 30–38, 2006.

[4] D. Wyatt, T. Choudhury, J. Bilmes, and H. Kautz, "A privacy-sensitive approach to modeling multi-person conversations," in *Proceedings of International Joint Conference on Artificial Intelligence*, 2007.

[5] D. Olguin-Olguin and A. Pentland, "Sensor-based organisational design and engineering," *International Journal of Organisational Design and Engineering*, vol. 1, pp. 5–28, 2010.

[6] E. Berke, T. Choudhury, S. Ali, and M. Rabbi, "Objective sensing of activity and sociability: Mobile sensing in the community," *Annals of Family Medicine*, vol. 9, pp. 344–350, 2011.

[7] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Proceedings of Workshop on Classification of Events, Activities, and Relationships and the Rich Transcription Meeting Recognition*, 2008.

[8] B. P. Milner and X. Shao, "Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction." *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 24 – 33, 2007.

[9] S. H. K. 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," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2010.

[10] J. Krumm, "A survey of computational location privacy," *Personal and Ubiquitous Computing*, vol. 13, pp. 391–399, August 2009.

[11] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, 2003.

[12] "http://www.nist.gov/speech/tests/rt/rt2007/spring/."

[13] R. Donovan, "Trainable speech synthesis," Ph.D. dissertation, Cambridge University, 1996.

[14] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of IEEE*, vol. 63, pp. 561–580, 1975.

[15] S. R. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, pp. 1243–1261, 2006.

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

[17] P. Thevenaz and H. Hugli, "Usefulness of the LPC- residue in text-independent speaker verification," *Speech Communication*, vol. 17, pp. 145–157, 1995.

[18] S. H. K. Parthasarathi, M. Magimai.-Doss, D. Gatica-Perez, and H. Bourlard, "Speaker change detection with privacy-preserving audio cues," in *Proceedings of International Conference on Multimodal Interfaces and Workshop on Machine Learning for Multimodal Interaction*, 2009.

[19] H. Bourlard and N. Morgan, *Connectionist Speech Recognition- A Hybrid Approach*. Kluwer Academic Publishers, 1994.

[20] H. Hermansky, D. Ellis, and S. Sharma, "Tandem Connectionist Feature Extraction for Conventional HMM Systems," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2000.

[21] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504 – 507, 2006.

[22] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, "Probabilistic and bottle-neck features for lvcsr of meetings," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2007.

[23] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring strategies for training deep neural networks," *Journal of Machine Learning Research*, vol. 10, pp. 1–40, 2009.

[24] J. Frankel, D. Wang, and S. King, "Growing bottleneck features for tandem ASR," in *Proceedings of Interspeech*, 2008.

[25] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proceedings of Advances in Neural Information Processing Systems*, 2006.

[26] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.0*. Cambridge University Press, 2000.

[27] S. Furui, "Research on individuality features in speech waves and automatic speaker recognition techniques." *Speech Communication*, vol. 5, pp. 183 – 197, 1986.

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

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

[30] F. Grezl and P. Fousek, "Optimizing bottle-neck features for LVCSR," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2008.

[31] J. Bridle, "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*, 1990.

[32] T. M. Cover and J. A. Thomas, *Elements of information theory*. Wiley, 1991.

[33] G. A. Miller, "Note on the bias of information estimates," *Information Theory and Psychology*, pp. 95–100, 1954.

[34] S. Chen. and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *DARPA speech recognition workshop*, 1998.

[35] "http://www.nist.gov/speech/tests/rt/rt2006/spring/."

[36] X. Anguera, "Beamformit, the fast and robust acoustic beamformer," in *http://www.icsi.berkeley.edu/x̃anguera/BeamformIt*, 2006.

[37] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, J. Vepa, and V. Wan, "The AMI meeting transcription system: progress and performance," in *Proceedings of Workshop on Machine Learning for Multimodal Interaction*, 2006.

[38] M. Sambur, "Selection of acoustic features for speaker identification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, pp. 176–182, 1975.

[39] K. S. R. Murty, B. Yegnanarayana, and S. Guruprasad, "Voice activity detection in degraded speech using excitation source information," in *Proceedings of Interspeech*, 2007.

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

[41] M. Wester, "The EMIME bilingual database," The University of Edinburgh, Tech. Rep., 2010.

[42] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez, "Estimating dominance in multi-party meetings using speaker diarization." *IEEE Transactions on Audio, Speech, and Language Processing*, 2011.