IDIAP Research Report

TRAP-TANDEM:
DATA-DRIVEN EXTRACTION OF TEMPORAL FEATURES FROM SPEECH

Hynek Hermansky

IDIAP-RR -03-50

31 AUGUST,
2003

Dalle Molle Institute
for Perceptual Artificial
Intelligence • P.O.Box 592 •
Martigny • Valais •

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

Abstract Conventional features in automatic recognition of speech describe instantaneous shape of a short-term spectrum of speech. The TRAP-TANDEM features describe likelihood of sub-word classes at a given time instant, derived from temporal trajectories of band-limited spectral densities in the vicinity of the given instant. The paper presents some rationale behind the data-driven TRAP-TANDEM approach, briefly describes the technique, points to relevant publications and summarizes results achieved so far.

1. Introduction

Machine that would automatically decode the linguistic information remains an elusive goal of engineering for many decades. Workers in automatic recognition of speech (ASR) face the similar challenge as human cognitive system does, i.e. to decode the information in the one-dimensional signal. Elaborate ASR systems, capable of acquiring and summarizing information contained in large amounts of training speech data, has been developed. However, existing ASR-based human-machine interfaces are still inadequate, fragile and unreliable in many realistic situations and environments encountered in human-human interactions. This prevents the wide acceptance of ASR technology by general public.

The problems starts at the very input to the machine, i.e. in the acoustic processing module. There, the speech signal is typically chopped into a short segments and the shape of the short-term spectral density is derived to yield data for the subsequent pattern classification. The pattern classification module then yields likelihoods of sub-word classes, used in the search for the best fitting hypothesis about the uttered sound sequence.

2. Human speech processing

ASR processing of speech signal is different from the way the speech signal is handled in human speech communication. Even though for steady sounds it may be possible to find some correlation between the shape of the sound spectrum and the level of activity on the auditory nerve, this correlation weakens when the sound intensity approaches levels encountered in speech communication [Sachs and Young 1979]. While there is no doubt that the auditory periphery is frequency selective, it is not clear that its main purpose is the deriving short-term spectrum of the acoustic signal. No natural sounds are steady but they change in time. It is desirable to find similar correlates of phonetic quality for naturally changing speech sounds, as the formants are in the steady-state vowels. It seems more likely that (consistently with color separation in vision) the selectivity of hearing is used for separating the reliable (high SNR) part of the signal from the unreliable ones. This is supported by the findings that for normal sound levels, temporal aspects of the sound need to be explored in order to account for the sound spectral shape in mammalian hearing [Young and Sachs 1979]. So our intuition is that parameters, which would account for phonetic quality of dynamic sounds, would be temporal.

3. Modulation aspect of information in speech signal

The relative fast changes in the acoustic pressure (20-20000 Hz) are merely the carrier of the acoustic information that is to be extracted from the signal. In human speech, the fast changes are caused by action of voice source (e.g. vocal cords in the case of voiced sounds). The slower modulations of the speech signal that carry the actual linguistic information, result from movements of vocal tract. Therefore the information that we are interested in machine recognition of speech is mostly encoded in the relatively slow modulations (below 50 Hz and likely not much higher than 10 Hz) of the acoustic wave [Kanedera et al. 1999]. Many perceptual phenomena such as forward masking, growth of loudness, detection of constant energy stimuli, or binaural release from masking, exhibit time constants of several hundreds of ms. Human hearing apparatus thus seem to poses the right kind of hardware for decoding the slow modulation changes. As discussed in the next paragraph, such time constants most likely originate at higher levels of neural systems.

4. Cortical receptive fields

The current knowledge about cortical responses to acoustic stimuli (cortical receptive fields) [e.g. de Charms et al 1998, Klein et al 2000, Depireux et al 2001] suggests that the auditory system is most likely to produce responses to certain time-and-frequency localized combinations of spectral densities in the time-frequency plane (acoustic events). One of such cortical fields (courtesy of David Klein) is shown in Fig. 1. It shows the spectro-temporal pattern of the auditory stimulus that is most likely to cause firing of the particular cortical neuron associated with this field. The neuron which would merely detect energy at the given time and frequency (e.g. the formant in speech) would have receptive field with a single high region at the given frequency and close to beginning of the temporal axis, the rest of the field would be close to zero. Such neurons do exist, but most cortical neurons have receptive fields far more complex than that. The length of a typical receptive field is up to several hundreds of ms, thus easily spanning the time span of dominant speech coarticulation. Both time and frequency resolution of the individual receptive fields varies rather widely with medians somewhere around 200 ms and 1 octave [Depireux et al 2001]. As discussed later, these relatively recent findings about physiology of auditory cortex may have important implications in ASR.

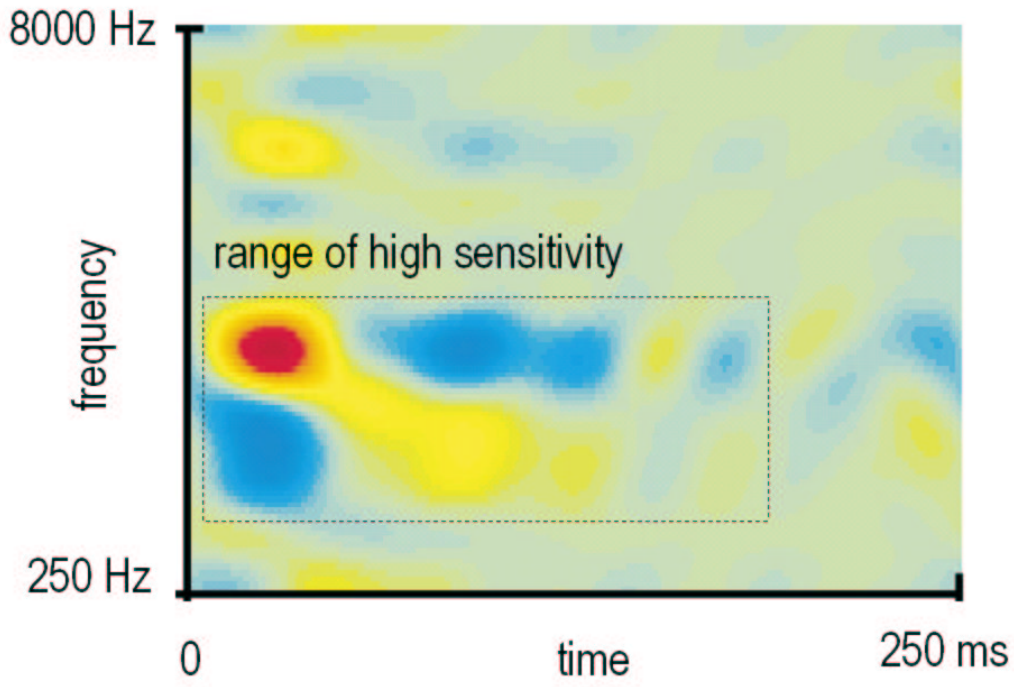

Figure 1 One of cortical receptive fields, observed in auditory cortex of ferret (courtesy of David Klein, used with permission, note in the figure by the author). Higher spectral density regions indicated by warm colors, lower by cold ones.

5. Current ASR

Limited knowledge of human speech communication process did not stop successful and profitable engineering applications of speech processing. In speech coding, the genius of inventors of telephony was in emulating the actions of the outer and middle ear and in converting the changes in the acoustic pressure into changes in electric current. The electric signal then could be transmitted and/or stored and used for reconstruction of the acoustic signal that closely resembles the original. Over the years, various techniques of digitizing and of efficient coding of the digitized electric signal evolved and are in daily use. The path of engineering attempts for recognition of speech followed that of speech coding. A typical first step in processing of speech signal in ASR is to convert it to a sequence of short-term spectral vectors, each vector describing frequency content of a single short segment of speech. The information-bearing temporal changes of the signal are then reflected in temporal changes of the short-term spectral vectors. This short-term spectrum of speech forms basis of features that are classified as belonging to different discrete information-carrying sub-word elements (states of the stochastic hidden Markov model). The information is decoded by finding the most likely path through the lattice of the discrete elements while respecting the prior knowledge about the possible distribution of the elements. Dynamics of speech is emulated by sequential organization of the elements.

6. ASR and human speech communication

Given the available knowledge about human auditory system, we would like to question some of the aspects of the current approach, and to suggest possible alternatives, which we consider to be more in line with our current knowledge of human hearing. Why should the knowledge of human hearing help in ASR? The human perceptual system appears to be optimally suited for decoding the information conveyed by sensory signals [Atick 1992, Malayath and Hermansky 2003]. Following human-like strategies in processing the cognitive signals is therefore a reasonable engineering way towards improvements of human-machine interface. Some of the aspects of hearing such as the nonlinear (critical-band like) frequency resolution or compressive nonlinearity between acoustic stimulus and its percept are well accepted by speech engineering community. Several times in the career of the author it happened that optimizing the ASR system resulted in processing that is consistent with human hearing. Some of this experience is summarized in the sections below.

Optimality of nonlinear frequency scale

ASR community is currently settled on two dominant and similar spectral processing techniques, the Mel cepstrum [Mermelstein 1976, Davis and Mermelstein 1980] and PLP [Hermansky 1990]. Both techniques employ auditory-like warping of short-term spectrum of speech, yielding higher spectral resolution at lower frequencies. The need for such non-uniform spectral resolution in ASR seems well established through years of comparative experiments. To what extent is this particular spectral warping optimal would typically require running a number of ASR experiments. Such optimization is costly and there is no guarantee that the solutions obtained will not be specific to a given ASR system. Therefore, we use a data-based optimization, which avoids using a specific ASR paradigm. Such a technique is based on the linear discriminant analysis (LDA), which is a stochastic technique that attempts to optimize the linear discriminability between classes in the presence of undesirable within-class variability (see e.g. [Hunt 1979, Brown 1987] for some examples of previous use of LDA in ASR).

LDA, applied to short-term spectral vectors from FFT analysis of OGI Stories database (OGI Stories contain about 3 hours of fluent American English telephone-quality speech from more than 200 adult speakers of both genders, hand-labeled by phonemes) yields spectral basis illustrated in Fig. 2. Notice that these spectral bases oscillate around zero faster at lower frequencies. Subsequently, speech analysis that employs such spectral basis has higher spectral resolution at lower frequencies. [Malayath and Hermansky 2002, Malayath and Hermansky 2003] show that the spectral resolution implied by spectral basis in Fig. 2 is very similar to spectral resolution of auditory-like Bark frequency scale. This finding supports earlier results of [Umesh et al. 1997] who derived auditory-like frequency warping by minimizing differences between speech from different talkers.

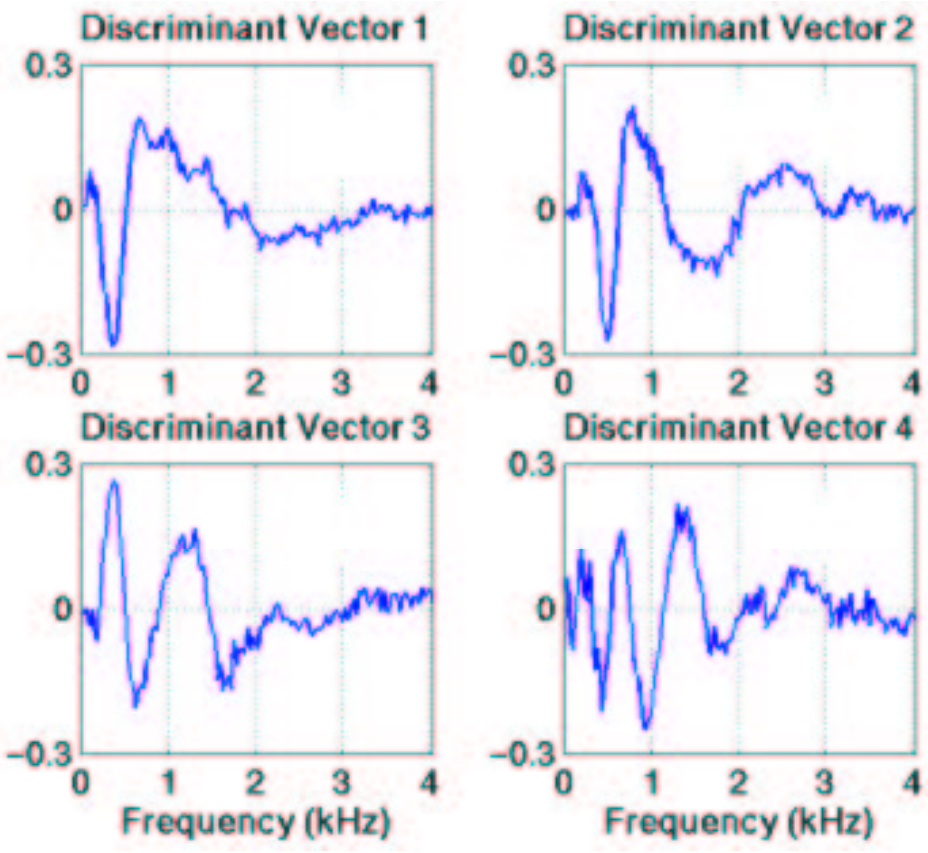

Figure 2 Spectral basis derived by data-driven LDA technique.

Data-guided design of RASTA-like filters

RASTA processing does filtering of time trajectories of speech features so that the features with rate-of-change that is not expected for speech, are attenuated. The initial ad hoc form of the RASTA filters [Hermansky and Morgan 1994] was optimized on a relatively small series of ASR experiments with noisy telephone digits. There is a way to structure the LDA problem in such a way that the LDA solution can be interpreted as a set of FIR RASTA-like filters, which are applied on time trajectories of spectral energies. This happens when the labeled vector space for LDA analysis is created by extracting temporal vectors cut out from trajectories of logarithmic critical-band spectral energy over a relatively long (typically about 1 s) span of time. Each vector typically spans much more than a single phoneme, and is labeled by the phoneme at the center of the vector.

Having formed such 101-dimensional (each vector spans about 1 s at 100 Hz sampling frequency) vector space with vectors labeled by their respective phoneme classes, LDA analysis yields a 101 X 101 scatter matrix, decomposed into its principal components. Then the principal vectors represent FIR filters, which most efficiently (with respect to the within-class and the across-class variability) map the 101-dimensional input space to several points of the output space.

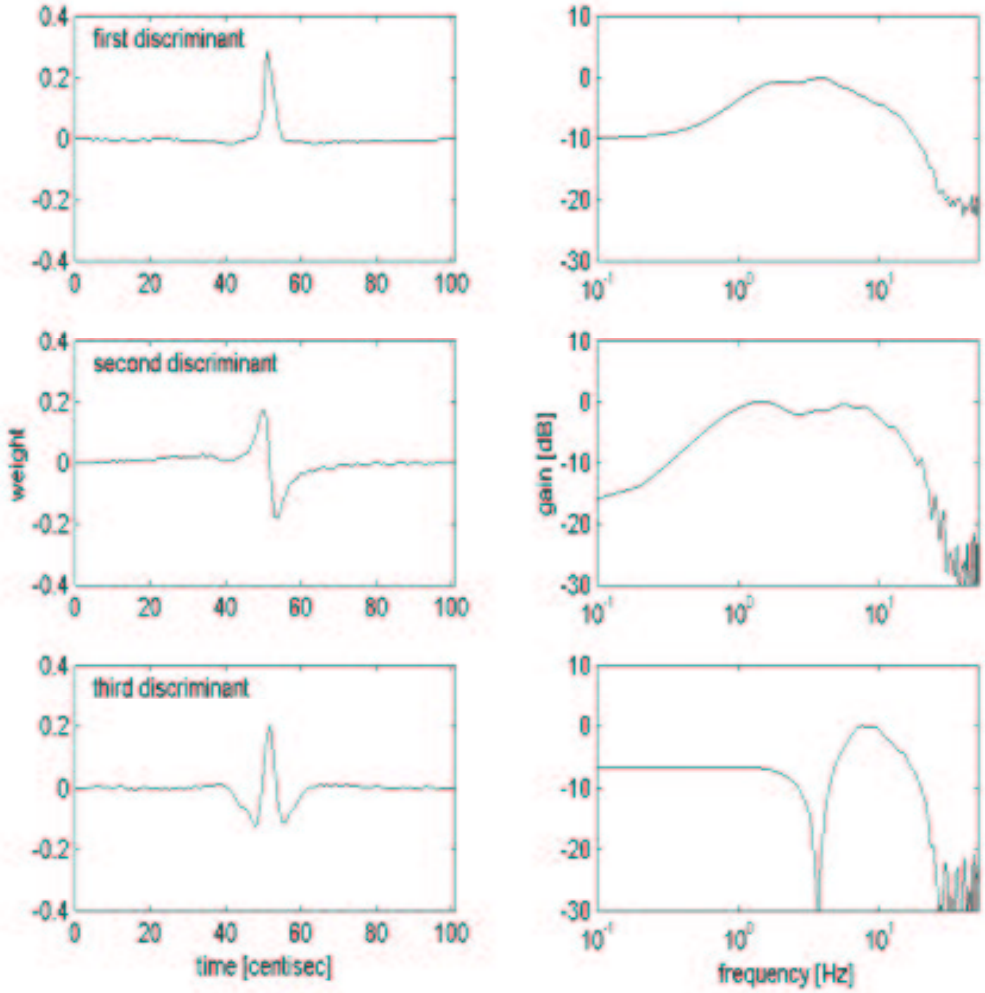

Figure 3 Impulse and frequency responses of the first three discriminant vectors from the LDA-derived discriminant matrix. The filters for the 5 Bark frequency channel are shown here. Filters for the other carrier frequencies studied (between 1 and 14 Bark) are very similar.

Frequency responses of the first three FIR filters derived from OGI Stories database are shown in Fig. A2. Filters for different frequency channels are similar. The frequency characteristic (shown at in the right part of the Figure) are generally consistent with RASTA [Hermansky and Morgan 1994], and delta, and double-delta feature of speech [Furui 1981]. However, impulse responses of the data-derived filters shown in the upper part of the figure suggest preference for the zero-phase filters. Effective parts of the impulse responses appear to span at least 250 ms.

The general characteristics of the data-derived RASTA filters appear to be relatively independent of the particular database used for their design. The most important processing involves a mild temporal lateral inhibition in which the average of several spectral values around the current time instant is subtracted from the weighted average of spectral values from surrounding past and future contexts. Next is the difference between weighted averages from

left and right contexts of the current frame (the first derivative of the first discriminant vector), followed by an aggressive mexican hat temporal lateral suppression (the second derivative of the first discriminant vector) implying quite narrow band-pass filter with 12dB/oct slope. Such dynamics-enhancing functions are hypothesized to be important for scene interpretation by human visual system [Marr 1982].

However, most of techniques employed in ASR are in many aspects inconsistent with hearing. We believe that improved understanding of the ways human perceptual system processes cognitive signals such as speech and images and of the methods of emulating such human-like processing by the machine would to necessary improvements of the human-machine communications.

7. Arguments against spectral envelope

Over the years, the concept of linear model of speech production and the emphasis on short-term spectral envelopes of speech, dominates the field. Emulating the temporal evolution of short-term spectral envelopes of speech could indeed lead to speech-like signals that conveys linguistic message of the original speech. Consequently, finding the spectral envelopes of speech forms basis of many speech coding techniques.

Since ASR evolved from speech coding, most of current ASR devices use stochastic pattern matching of features, which are derived from short-term spectral envelopes of speech sounds. The short-term spectral envelope is usually modified by nonlinear warping of its frequency (Mel or Bark scale) and amplitude axes (logarithm) and projected on spectral basis that decorrelate the feature space (cepstrum). However, that does not change the fact that the estimate of phonetic quality of incoming speech segment is based on the shape of the spectral envelope.

However, coding of linguistic information in a single short-term spectrum of speech appears to be rather complex. A single frame of short term spectrum does not contain all the information that is necessary for decoding the phonetic value of a given segment of speech. This is because the neighboring speech sounds influence the short-term spectrum of the current sound. The mechanical inertia of human speech production organs (coarticulation) results is significant spreading of linguistic information in time (our current estimate would be of the order of several hundreds of ms [Yang et al 2000]). Given the typical phoneme rate at about 15 phonemes per second, this means that at any given time, at least 3-5 phonemes interact. This introduces high within-phoneme variability of the instantaneous spectral envelope. Some studies indicate that the within class variability is comparable in magnitude to the across-class variability among phoneme classes [Kajarekar and Hermansky 2000].

ASR attempts to classify phonemes from individual slices of the short-term spectrum and needs to deal with this within-class variability. This is often done by increasing number of sounds to be classified, i.e. by introducing so called context-dependent phonemes and by sub-dividing phonemes into several parts, each of which is emulated by a separate model. Both techniques lead to more complex ASR models. However, human listeners appear to be able to identify phonemes independently of their context [Fletcher 1953] in spite of large variability introduced

by their phonetic environment. This observation suggests that the highly-variable spectral envelope may play less role in human speech perception that it does in current ASR.

8. Cortical receptive fields, events in speech, and speech recognition.

As described earlier, neurons in the auditory cortex best respond to certain kinds of acoustic signals (e.g. [Klein et al 2000]). They seem to act as a kinds of two-dimensional matched filters, which could detect the existence of particular patterns in the incoming signal. Then, a certain combination of particular patterns could indicate certain sound class such as particular phoneme of the language.

This picture is not as far fetched from our current thinking about speech as it may look. Most would agree that an important (and well accepted) model of speech communication uses a concept of formants that are represented in the short-term spectral frame as peaks of the short-term spectral envelope. Think about a vowel formant as one particular type of an acoustic event, characterized by a rather trivial time-frequency localized pattern, consisting of high vocalic energy at the given time instant and at frequencies in the neighborhood of the formant frequency. Such a receptive filed is shown in the left part of Fig. 4.

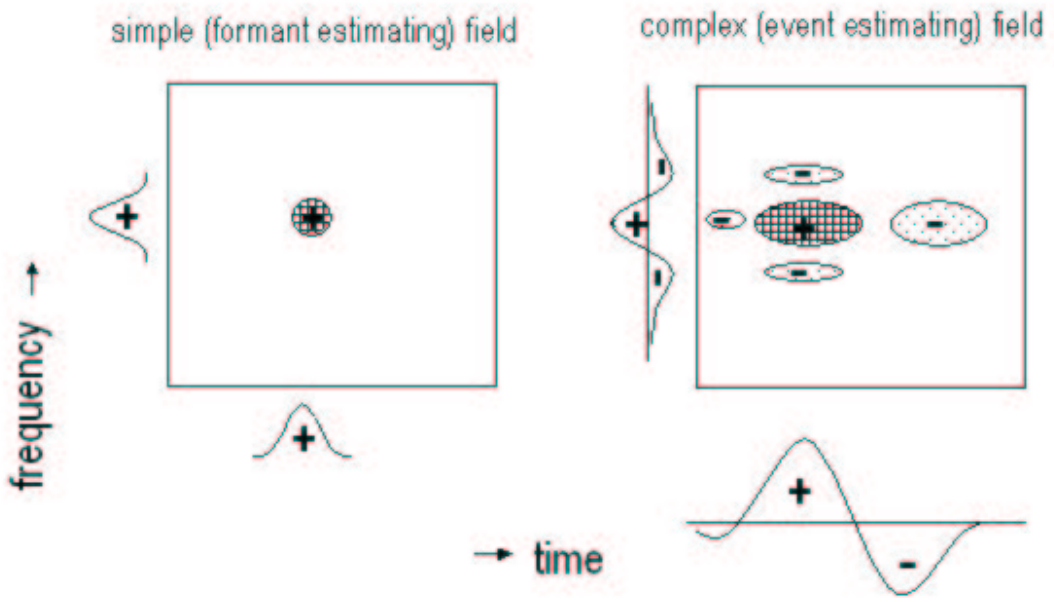

Figure 4 Schematic examples of simple and more complex receptive fields

Since the typical cortical receptive field is more complex than a having only a single short excitatory region, (a schematic example of such more complex receptive field is depicted in the

right part of Fig. 2), we are merely extending the notion of the formant to more complex events. As described earlier, the cortical receptive fields span up to several hundreds of ms and up to several octaves and exhibits not only excitatory but also inhibitory regions. Cortical neurons associated with such receptive fields would optimally respond to more complex acoustic events than the steady formants of speech. Such broadly defined events can be than characterized by more complex time-frequency patterns, possibly involving particular combinations of high and low spectral energies at times other than the current instant.

9. TRAP-TANDEM

Introduction to the technique

Now, how do we use our new notion of complex time-frequency acoustic events in an automatic speech recognizer? First, we need to realize the needs of state-of-the-art stochastic recognizers. Ideally, the ASR system expects feature vector of (within state) uncorrelated and Normally distributed features every 10 ms or so. Further, the feature vectors should me small in size so that the subsequent pattern classifier is also small and could be trained on a finite amount of training data. Smallest set of features for classification are posterior probabilities of the classes to be classified [8]. So we need a module that would be capable of examination of relatively long spans of speech signal within various frequency bands, and to deliver every 10 ms or so posterior probabilities of particular temporal events within such bands and to convert these posteriors to a small set of uncorrelated and Normally distributed features.

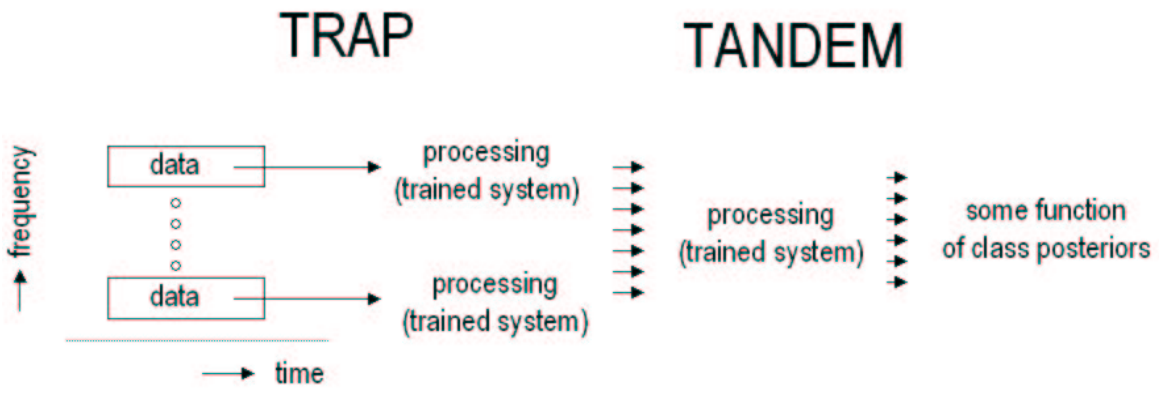

Figure 5 TRAP-TANDEM feature extraction

A step in this direction is the TRAP-TANDEM and related techniques [Hermansky and Sharma 1999, Hermansky et all 2000]. Schematic picture of the TRAP-TANDEM techniques is shown in Fig.5. The TRAP refers to a particular way in which the linguistic information is extracted from the speech data. In a conventional speech analysis, the spectral shape of full-band

spectrum a short segment (about 10-20 ms) of speech signal is used to provide evidence for the subsequent stochastic recognition techniques. In TRAP, the multiple evidence is derived from a relatively long (500-1000 ms) and frequency-localized (1-3 Bark) overlapping time-frequency regions of the signal. The TANDEM refers to a way of converting the frequency-localized evidence to features for the HMM-based ASR system. Both the TRAP and the TANDEM modules are trained on development data.

Why long time-spans of the signal?

Why would we attempt to derive speech features from time intervals as long as 1 s? Because the information about the underlying sub-word classes (phonemes) spreads at least over the interval of 200-300 ms. This has been demonstrated by Bilmes [Bilmes 1998] and confirmed by Yang et al. [Yang et al. 2000]. Since the derived features will be used for classification into phoneme-like classes, it makes sense to collect the evidence from all the data points which carry the information, hence at least 300 ms. But why even longer time interval? Because we want to remove the information about slowly varying noise (subtract the mean) from the data. This harmful information is in modulation spectrum below 1 Hz [Kanedera et al 1999], hence 1 s.

Why independent frequency-localized processing?

Why would we abandon the short-term spectrum of speech? First, as already mentioned, the envelope of the short-term spectrum is notoriously unreliable in presence of common distortions such as the distortions caused by frequency response of communication equipment or by frequency localized noise. Fletcher [Fletcher 1953] (and many after him) demonstrated that uncorrelated noise outside the critical band has only a negligible effect on detection of the signal within the critical band. He further proposes that errors in human recognition of nonsense syllables within relatively narrow articulatory spectral bands (each articulatory band spanning about 2 critical bands) are independent. Hence, the first stage of processing of acoustic signals seems to happen on frequency-localized regions of the signal.

Why training of the feature extraction module?

Why do we need to train the analysis module to derive features that will then be used in another trained stochastic system? Because more knowledge we build into the feature extraction module, less we need to train the subsequent stochastic recognizer. Our knowledge about coding of information in speech signal is still incomplete. As evidenced by the success of data-driven stochastic pattern classification and language modeling methods (see e.g. [Jelinek 1998] for details), using the incorrect prior knowledge may be worse than using no prior knowledge at all, and rather to derive all the required knowledge *a posteriori* from the data. Enough speech data, labeled with respect to the targeted linguistic message (either by hand or by forced alignment procedures) is available. It makes sense to use this data to train the feature extraction module and to derive speech-specific and task-independent knowledge for the recognizer.

Details of the technique

TRAP

The time-frequency spectral density plane currently being estimated using the front end module from PLP analysis [10]. It employs the short-time spectral analysis of the speech signal with a subsequent Bark-like summation of the spectral components. However, recently emerging interesting alternative for estimating temporal evolution of critical band spectral density that completely eliminates the short-term spectral analysis is the frequency domain linear prediction [Athineos and Ellis 2003].

The input to the TRAP estimator is formed from 1-3 time trajectories of critical-band energies. These temporal patterns may be parameterized. The most successful parameterization has been so far the cosine transform (closely resembling PCA (Karhunen-Loeve) transform (KLT) derived on development data). The cosine transform typically allows for at least 50% reduction of dimensionality of the input data. Some benefits are seen when more than one time trajectory is used as an input. In that case, the individual trajectories are concatenated to form a longer input vector, and some form of dimensionality reduction is typically applied. The PCA analysis of the data suggests in this case, that for preserving most of the variability in the multiple-trajectory data, the data from the individual trajectories should be averaged and differentiated, in effect crudely describing the spectral shape in the vicinity of the frequency of interest [Jain and Hermansky 2003, Grezl and Hermansky 2003]. The TRAP technique is depicted in Fig. 6. Fig. 7 schematically shows a few principal bases, resulting from the PCA analysis on three concatenated trajectories of critical band spectral densities. As seen in the figure, the lower KLT bases represent cosine transform on averaged temporal trajectories, some higher ones represent cosine transform on differentiated trajectories. It is interesting, that even though the differentiating bases account for relatively little variability, their elimination results in noticeable worsening of the performance [Jain and Hermansky 2003].

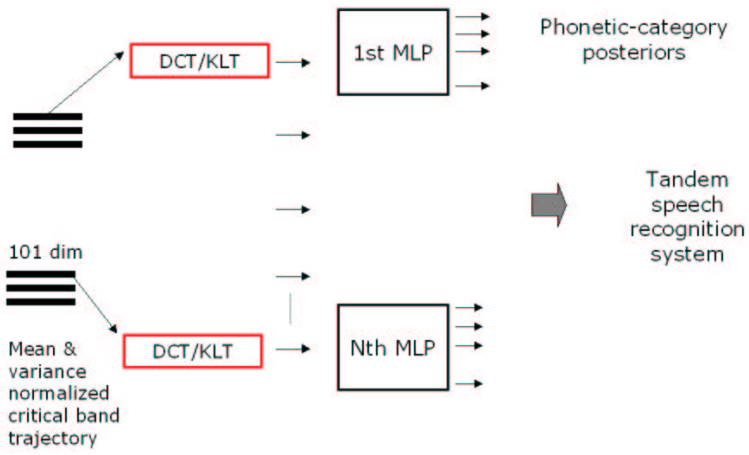

Figure 6 TRAP module

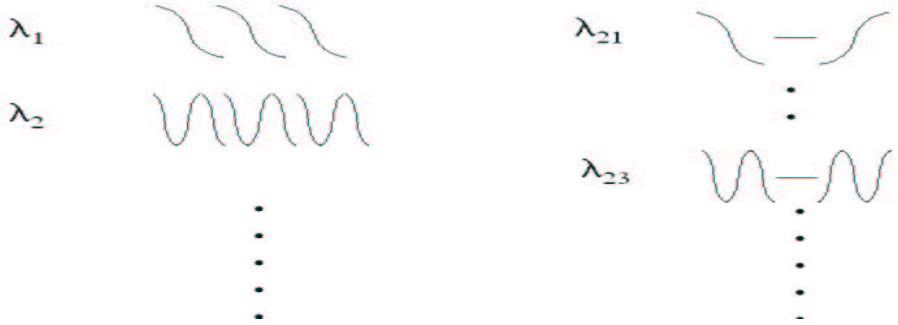

Figure 7 Schematic illustration of basis functions of the KLT transform derived for the three-band TRAP input

TRAP estimators deliver vectors of posterior probabilities of sub-word acoustic events, each estimated at the particular individual frequency. The events targeted by TRAP estimators are most often American English phonemes clustered into 6 broad phonetic classes [Adami et al 2002, Jain 2003] and separate estimator is being trained for each frequency region of interest. More recently, there are efforts to derive a single "universal" estimator, which could be used at all frequencies of interest [Hermansky and Jain 2003]. This would be much more in line with our "event" concept outlined above.

TANDEM

The TANDEM part of the technique derives a vector of posterior probabilities of sub-word speech events for every speech analysis frame from the evidence presented to its input. Techniques based on optimal rotation of feature space such as linear discriminant analysis (LDA) has been used in feature extraction in ASR for quite some time [Hunt 1979]. Nonlinear alternative to LDA is a multi-layer Perceptron (MLP) trained in one-high, rest-low paradigm. When properly trained, such MLP estimates posterior probabilities of classes of interest [Bourlard and Wellekens 1990, Bourlard and Morgan 1994].

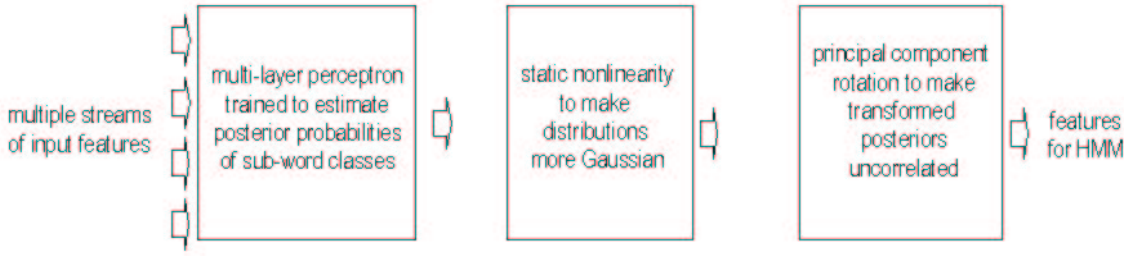

Figure 8 TANDEM technique for deriving features for HMM-based ASR

The MLP posterior probability estimates are gaussianized by a static nonlinearity and whitened by the KL transform derived from training data. Such gaussianized and whitened posterior probabilities form the feature vector for the subsequent HMM recognizer. Thus, we are replacing the conventional features derived from a spectral density vector representing the spectral envelope, by a matrix of transformed likelihoods of acoustic events (in the original concept the events were context-independent phonemes). If the targeted events are independent, the output of the trained TANDEM MLP could represent an estimate of the efficient low-entropy statistically-independent code, hypothesized in perceptual processing [Atick 1992, Lewicki 2002].

The events targeted by the TRAP estimators may be but do not need to be the same at the events targeted by the TANDEM estimator. At the moment, the events targeted by the TRAP are smaller in size (broad phonetic classes) where the targets for the TANDEM estimator are typically context-independent American English phonemes. Also, TRAP estimators can be (and often are) trained on different database than the database used in training the TANDEM estimator. Both the TRAP and the TANDEM estimators are nonlinear feed-forward Multi-Layer Perceptron (Quicknet [www.icsi.Berkeley.edu/speech/ICSI_SPEECH_FAQ]) discriminative classifiers. Hierarchical classification schemes in TANDEM estimator were also investigated [Sivadas and Hermansky 2002].

Results

The TRAP-TANDEM ASR has been so far found most useful in combination with the conventional spectrum-based (PLP, Mel Cepstrum,..) ASR. Thus, e.g. the system with TRAP-TANDEM module was shown to perform the best among all presented feature extraction techniques (including the officially accepted ETSI standard) on the small vocabulary Aurora task [Adami et al 2002]. More recently, the TRAP-TANDEM features were successfully used in DARPA EARS program, where they brought more than 10% relative improvement in error rate on a smaller (500 word task) and scaled successfully on a full vocabulary task [Morgan 2003].

However, the TRAP-TANDEM features are already becoming competitive with the traditional approaches. E.g., on the small vocabulary continuous OGI Numbers task, the three-band TRAP-TANDEM system yields the same (5%) word error rate as the best system with the conventional (PLP+delta+ddelta) features [Jain 2003]. In phoneme-string recognition without use of any language model on TIMIT database, using TRAP features in the PLP-HMM hybrid system gave about 10% relative improvement in phoneme error rate, comparing to the best multiframe Mel cepstral features [Schwarz et al 2003].

10. Discussion and conclusions

In several aspects, TRAP-TANDEM represents a significant conceptual departure from the current practice in feature extraction for ASR.

- The knowledge used for feature extraction is not all coming from beliefs and convictions of the designer but is mostly derived from development data. The goal here is to derive and to put into the feature extraction module the speech-specific but task-independent

knowledge. In that way, the subsequent pattern classification module would need to learn only the task-specific knowledge, possibly reducing the need for the re-learning the same knowledge again and again each time when the task changes.

- Derived features do not represent shape of the short-term spectral envelope of speech. Instead, in the early stages of the feature extraction, the frequency-localized evidence is converted to frequency-localized estimates of likelihoods of speech events (the TRAP part). These estimates are then used in later stages of the feature extraction (the TANDEM part). In that way, many vulnerabilities of the short-term spectral envelope of speech (discussed earlier in this paper) are alleviated.

- Evidence used for deriving the features does not all come from the relatively short segment of speech representing a short part of the underlying sub-word class (phoneme) but the employed time span covers at least the typical coarticulation span of the phoneme. In that way, each feature vector could carry most of the available information about the underlying phoneme.

- Final features represent estimates of posterior probabilities of sub-word classes postulated in the subsequent HMM-based pattern classification. In that way, the feature set could be smaller and the burden on the subsequent HMM classifier could be reduced.

The TANDEM-TRAP technique is still evolving and in order to get most out of it, it may also require some evolution of the rest of existing dominant ASR approach. However, it already yields useful supplementary information for the existing mainstream HMM-based ASR and is becoming competitive on its own. Unlike the conventional feature extraction approaches, it is consistent with the current knowledge of higher cognitive levels of mammalian auditory perception. We hope that it receives critical attention of the ASR community.

Acknowledgements

The insights presented in this paper come from research of many of my colleagues, most of them hopefully acknowledged through references to their works. Works of my former and current students Sangita Sharma, Pratibha Jain, and Sunil Sivadas are particularly relevant. The work was supported by DARPA EARS program and by grant from Qualcomm Inc. Writing of the paper was supported by Swiss National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2). The NCCR is managed by Swiss National Science Foundation on behalf of the Federal Authorities.

References

A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan and S. Sivadas (2002), "QUALCOMM-ICSI-OGI Features for ASR", in Proceedings ICSLP 2002, Denver, Colorado, USA, Sep, 2002

M. Athineos and D. Ellis (2003), “Frequency-domain linear prediction for temporal features”, Proc. IEEE ASRU-2003 Workshop, St. Thomas, US Virgin Islands, 2003

J.J. Atick (1992), “Could information theory provide an ecological theory of sensory processing?”, in Network: Computation in Neural Systems, Vol. 3, pp. 213-251, 1992

J. Bilmes (1998), “Maximal mutual information based reduction strategies for cross-correlation based joint distributional modeling”, Proc. ICASSP98, SP14.6, Seattle, 1998

P. Brown (1987), The Acoustic-Modeling Problem in Automatic Speech Recognition, PhD Thesis, Computer Science Department, Carnegie Mellon University.

H. Bourlard and Ch. Wellekens (1990), “Links Between Markov Models and Multilayer Perceptrons.”, IEEE Transactions on Pattern Analysis and Machine Intelligence 12: 1167-1178, 1990

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

D.D. Depireux, J.Z. Simon, D.J. Klein, S.S. Shamma (2001), “Spectro-Temporal Response Fields Characterization with Dynamic Ripples in Ferret Primary Auditory Cortex”, in J. Neurophysiology, Vol. 85, pp. 1220-1234, 2001

S.B. Davis, and P. Mermelstein (1980), Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans

C.R. deCharms, D. Blake, M.M. Merzenich (1998), Optimizing sound feature for cortical neurons, Science, Vol. 280, May 29, 1998

H. Fletcher (1953), *Speech and hearing in communication*, The ASA edition, edited by J.B. Allen, Acoust. Soc. Am., reissue of the original edition from 1953

K. Fukunaga (1990), *Statistical Pattern Recognition*, Academic Press, San Diego, 1990.

S. Furui (1981), Cepstral analysis technique for automatic speaker verification, IEEE Trans. on Acoustic, Speech, & Signal Processing, vol. 29, pp.254-272.

F. Grezl and H. Hermansky (2003), “Local averaging and differentiating of spectral plane for TRAP-based ASR”, Proc. Eurospeech 2003, Geneva 2003

H. Hermansky (1990), Perceptual linear predictive (PLP) analysis of speech, J. Acoust. Soc. Am., vol. 87, no. 4, pp. 1738-1752.

H. Hermansky (1998), "The Modulation Spectrum in Automatic Recognition of Speech", in 1997 IEEE Workshop on Automatic H. Hermansky and S. Sharma, "TRAPS Classifiers of Temporal Patterns", in ICSLP'98, Sydney, Australia, 1998,

H. Hermansky and P. Jain (2003),”Band-independent speech event categories for TRAP based ASR”, Proc. Eurospeech 2003, Geneva 2003

H. Hermansky (1998), "Should recognizers have ears?", in Speech Communication, vol. 25, num. 3-27, 1998

H. Hermansky and D.P.W. Ellis and S. Sharma (2000), "Connectionist Feature Extraction for Conventional HMM Systems", in Proc. ICASSP'00, Istanbul, Turkey, 2000

H. Hermansky and S. Sharma (1999), "Temporal Patterns (TRAPS) in ASR of Noisy Speech", in Proc. ICASSP'99, Phoenix, Arizona, USA, Mar, 1999

H. Hermansky and N. Malayath (1998) "Spectral Basis Functions from Discriminant Analysis", in Proc. ICSLP'98, Sydney, Australia, 1998

M.J. Hunt (1979), A statistical approach to metrics for word and syllable recognition, J. Acoust. Soc. Am., 66(S1), S35(A).

www.icsi.berkeley.edu/speech/faq/ICSI_SPEECH_FAQ

P. Jain (2003) PhD. thesis, Department of Electrical and Computer Engineering, OGI School of Oregon Health & Sciences University, Portland, Oregon, 2003

P. Jain and H. Hermansky (2003), "Effect of combining temporal patterns from critical-bands on ASR ", Proc. Eurospeech 2003, Geneva 2003

F. Jelinek (1998), *Statistical Methods for Speech Recognition*, MIT Press, 1998

D.J. Klein, D.A. Depireux, J.Z. Simon, S.S. Shamma (2000)," Robust spectro-temporal reverse correlation for auditory system: Optimizing stimulus design", in J. Comp. Neuroscience, Vol. 9, pp. 85-111, 2000

S. Kajarekar, and H. Hermansky (2000), " Analysis of information in speech and its application in speech recognition", in Proceedings of Workshop in Text, Speech and Dialogue 2000, Brno, Czech Republic, Springer-Verlag.

N. Kanedera, T. Arai, H. Hermansky and M. Pavel (1999), "On the relative importance of various components of modulation spectrum for automatic speech recognition", Speech Communication, 28, (43-55), Elsevier 1999

M.S. Lewicki (2002), "Efficient coding of natural sounds", Nature Neuroscience, 5(4), pp. 356-363, 2002

N. Malayath and H. Hermansky (2002), Bark resolution from speech data, Proceedings International Conference on Spoken Language Processing 2002, Denver, Colorado, September 2002.

N. Malayath and H. Hermansky (2003), Data-driven spectral basis functions for automatic speech recognition, Speech Communication, Vol. 40 (4), pp. 446-466, June 2003.

N. Morgan et al. (2003), DARPA-EARS Meeting, Boston, MA, May 2003

D. Marr (1982), Vision, W.H. Freeman, San Francisco.

P. Mermelstein (1976), Distance measures for speech recognition, psychological and instrumental, in Pattern Recognition and Artificial Intelligence, R.C.H. Chen, ed., Academic Press: New York, pp. 374-388.

M. Sachs and E. Young (1979), “Encoding of steady state vowels in the auditory nerve: representation in terms of discharge rate”, J. Acoust. Soc. Am. 66, pp. 470-479, 1979

P. Schwarz, P. Matejka and J.Cernocký (2003), “Recognition of Phoneme Strings using TRAP Technique”, Proc. Eurospeech 2003, Geneva 2003

S. Sivadas and H. Hermansky (2002), "Hierarchical Tandem Feature Extraction", in Proceedings ICASSP 2002, Orlando, Florida, USA, May, 2002

S. Umesh, L. Cohen and D. Nelson (1987), Frequency warping and speaker normalization, Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Munich, Germany, 1997, pp. 983-987.

H. H. Yang, S. Sharma, S. van Vuuren and H. Hermansky (2000), "Relevance of Time-Frequency Features for Phonetic and Speaker/Channel Classification", in Speech Communication, Aug, 2000

E. Young and M. Sachs (1979), “Representation of steady-state vowels in the temporal aspects of the discharge patterns of population of auditory nerve fibers, J. Acoust. Soc. Am 66, pp. 1381-1403, 1979.