Detection and Identification of Rare Audiovisual Cues

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect)   Plautus (ca 200(B.C.)

Project no: 027787

DIRAC

Detection and Identification of Rare Audio-visual Cues

Integrated Project
IST - Priority 2

DELIVERABLE NO: D1.3
Acoustic Features that Employ Limited Frequency Ranges and Longer Temporal Spans

Date of deliverable:31.12.2006
Actual submission date:31.01.2007

Start date of project: 01.01.2006 Duration: 60 months

Organization name of lead contractor for this deliverable: **IDIAP Research Institute**

Revision [1]

Project co-funded by the European Commission within the Sixth Framework Program (2002-2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other program participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

DIRAC
Detection and Identification of Rare Audiovisual Cues

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than
things you do expect)   Plautus (ca 200(B.C.)

D1.3 Acoustic Features that Employ Limited Frequency Ranges and Longer Temporal Spans

IDIAP Research Institute (IDIAP)

Abstract:

Conventional features in automatic recognition of speech describe the instantaneous overall shape of a short-term spectrum of speech. In this report we summarize results of ongoing research towards alternative speech features that rely on information in temporal dynamics of spectral energy. We take an extreme position by ignoring any long-distance correlations among spectral components of the short-term spectrum of speech, and considering temporal trajectories of spectral energies as carriers of information in the acoustic signal. This approach is inspired by observed properties of auditory cortical receptive fields and supported by results of data-driven feature extraction based on Linear Discriminant Analysis. A technique for all-pole autoregressive modelling of Hilbert envelopes in sub-bands is investigated as means for modelling the temporal trajectories. Finally, the long temporal trajectories of spectral energies in critical-bands are used as features in modulation spectrum based module for discrimination of speech carrying segments of the signal. The feasibility of the investigated novel features is supported by results on well established speech recognition tasks.

Table of Content

1. Introduction

1.1. Speech

Sounds represent important interface with the outside world. In the form of speech, they provide for one of the most important cognitive functions, for the language communication. Speech is formed by a sequence of particular sounds. Under certain conditions, these sounds could be called phonemes of the language. Sequences of phonemes form words, sequences of words form phrases, and the phrases may carry the linguistic message in speech. Thus, how do particular vibrations of the molecules of air create the percept of speech sounds in the human hearing system is a matter of great interest.

Most of accepted concepts of speech perception start with a spectral content of phonemes. Indeed, different speech sounds are characterized by different spectral contents. From the speech production point of view it makes sense. Human vocal tract resonates at certain frequencies, called formants of speech. It is relatively straightforward to modify resonance properties of human vocal tract by changing its shape. From the perceptual point of view, the spectrum as a carrier of the information in the speech signal is supported by a spectral selectivity of early stages of mammalian auditory perception.

Speech is formed by sequences of sounds, and these sounds originate in changes in vocal organs with their finite inertia. Therefore, spectral content of the speech signal continuously and gradually changes to convey the information. These changes can be seen in the spectrogram, which represents sequence of short-term spectra, each computed from a very short segment of the speech signal. Since the changes are gradual, the short-term spectra change within the phoneme boundaries. Further, since the inertia of speech production organs spans more than the duration of a typical phoneme, spectra within the phoneme are influenced by neighboring phonemes (coarticulation). In addition, the spectrum of speech depends on inherent anatomy of a particular speaker. As a consequence of all these facts, speech sounds that represent identical phonemes may be represented by different short-term spectral vectors. Finally, it is easy to change many aspect of the short-term speech spectrum by common disruptions that would have hardly any effect on human speech communication. All this makes us to look for alternatives to short-term spectrum of speech.

In spite of that, human listeners are able to decode individual and unique phonetic elements from the speech stream. E.g., highly trained listeners in extensive experiments with perception of meaningless syllables done at Bell Laboratories in the first half of the last century were reported to perform in average with better than 98 % accuracy when transcribing the individual phonemes in the syllables [12]. How is this done is a matter of great theoretical and practical interest. Theoretical, since it would contribute to understanding of the unique process of human speech communication and could have important implications for our understanding of processing of cognitive signals in general. Practical, since it may allow for more accurate machine decoding of the information in speech.

1.2. Modulation Spectrum of Speech

One viable alternative, proposed and being studied at the I.P. Pavlov Institute in Leningrad/St Petersburg in the former Soviet Union in sixties and seventies of the last century, is to evaluate not only the instantaneous short-term energy of the spectrum at various frequencies but rather to derive frequency components of temporal trajectories of spectral energies in the vicinity in a given time instant. Such so called modulation spectrum of speech then caries information about spectral dynamics in the vicinity of a given time instant, thus providing the information not available in the short-term spectral energy (which represents a mere DC component of the modulation spectrum). Further, some components of the modulation spectrum may be less affected by certain common distortions of the signal. This has been applied with advantage in techniques such as RASTA processing [24] that operates as a bandpass filter on the modulation spectrum of speech, preserving only its components that are presumed to carry the speech information.

1.3. Speech Technologies

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. Most of efficient speech coding techniques first convert speech signal to a sequence of short-term spectral vectors, each vector describing frequency content of a single short segment of speech. Aspects of these short-term spectral vectors are extracted and transmitted to the receiver, where the speech is reconstructed.

However, machine that automatically decodes the linguistic information still remains an elusive engineering goal. Workers in automatic recognition of speech (ASR) face a similar challenge as human cognitive system does, i.e. to decode the information in the one-dimensional signal. In spite of that, ASR processing of a speech signal is different from the way the speech signal is handled in human speech communication. This is partly because many aspects of human information processing are becoming to be known only recently and many are still shrouded in mystery.

Since ASR evolved from speech coding, the feature extraction module in ASR typically resembles techniques from speech coding. The speech signal is first chopped into a short segments and the shape of the short-term spectral density is derived to yield data for the subsequent pattern classification. Thus, the one-dimensional speech signal is converted into a sequence of short-term spectral vectors, each vector describing frequency content of a single short segment of speech. This short-term spectral vector is typically transformed by series of ad-hoc transformations into the feature vector for the subsequent pattern classification. The local spectral dynamics in included by use of the first (delta) and the second (delta-delta) temporal derivatives of the feature trajectory in the vicinity of the current element [14].

In the currently dominant stochastic ASR, the sequential pattern classification module is applied to decode the information in feature vectors. This is done by first obtaining likelihoods of sub-word elements (states of the stochastic hidden Markov model), used in the search for the best fitting hypothesis about the uttered sound sequence. The information is decoded by finding the most likely path through the lattice of these discrete elements while respecting the prior knowledge about the possible distribution of the elements. The global dynamics of speech is emulated by sequential organization of the elements.

The pattern classification module is relying on information extracted from large amounts of acoustic and text training data. Elaborate ASR systems, capable of acquiring and summarizing the information contained in large amounts of training data, have been developed. Still, existing ASR-based human-machine interfaces are 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.

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. Given the knowledge about human auditory system, we would like to question some of the aspects of the current approach, and to suggest possible alternatives, more in line with the 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 [3, 36]. Following human-like strategies in processing the cognitive signals is therefore a reasonable engineering approach 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.

1.4. Short Spectrum of Speech

An important (and well accepted) model of human speech communication uses a concept of resonance frequencies of the vocal tract (formants). The formants show in the short-term spectrum as peaks of the short-term spectral envelope. Accurate emulation of time-varying formants yields intelligible speech [30, 8]. Over the years, the concept of a linear model of speech production and the emphasis on short-term spectral envelopes of speech dominate the field and finding the spectral envelopes of speech forms basis of many speech coding techniques.

As discussed above, most 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 cosine spectral basis (computation of so-called cepstrum) that decorrelate the feature space.

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 estimates are of the order of several hundreds of ms [44]). Given the typical phoneme rate of about 15 phonemes per second, this means that at any given time, at least 3-5 phonemes interact. Some studies indicate that the within class variability is comparable in magnitude to the across-class variability among phoneme classes [32]. The coarticulation effects, combined with many additional sources of nonlinguistic variability such as speaker identity and the effects of the acoustic environment, all contribute to high within-phoneme variability of the instantaneous spectral envelope. Subsequently, coding of linguistic information in a single short-term spectrum of speech appears to be rather complex. Indeed, it has been suggested that in order to derive phoneme identity from the running speech, one needs to collect the information from the whole speech syllable (i.e. from about 200-300 ms of the signal) [52].

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, e.g. 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 [12] in spite of large variability introduced by the coarticulation with the neighboring phonemes. This observation suggests that the coarticulation effects, while clearly evident in temporal evolution of spectral envelopes, may not present the same problem in human speech perception that they do in current ASR.

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. 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 [40]. 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 [45].

ASR community is currently settled on two dominant and similar spectral processing techniques, the Mel cepstrum [38, 10] and PLP [16]. 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.

1.5. Temporal Aspect of Speech Signal

The relatively 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 [33]. Clearly, to obtain sufficiently low frequency components of the modulation spectrum of speech, a sufficiently long segment of spectro-temporal plane is necessary. 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 milliseconds. As discussed later in this article, such time constants most likely originate at higher levels of neural systems. Thus, the human hearing apparatus seems to have right properties for decoding the slow modulation changes.

The temporal aspects of speech have been stressed and extensively studied in works of several speech groups, probably nowhere more than in the Chistovich-Kozhevnikov group
In the former Leningrad (now again St Petersburg) in Russia. Their conclusions, unfortunately still only rather speculative and partly reported in their 1965 book that was translated to English [52] but much more convincingly and with more experimental support reported in the later 1976 book [56] available only in Russian, are very specific: “ Human listeners are on one hand able to extract parts of the speech signal corresponding to the individual speech sounds but on the other hand they need to use the information present in the neighboring phonemes in the process” ([56], p. 68).

1.6. Spectro-Temporal Aspect of Mammalian Auditory Processing

Hearing in mammals possesses a mechanism for processing longer segments of the signal. The current knowledge about cortical responses to acoustic stimuli (cortical receptive fields) [11, 31, 9] suggests that the auditory system is most likely to produce responses to specific time-and-frequency localized combinations of spectral densities in the time-frequency plane (acoustic events). One cortical field (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. The neuron merely detecting 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 milliseconds, thus easily spanning the time span of speech co articulation. Both the time and the frequency resolution of the individual receptive fields vary rather widely with medians somewhere around 200 ms and 1 octave [9].

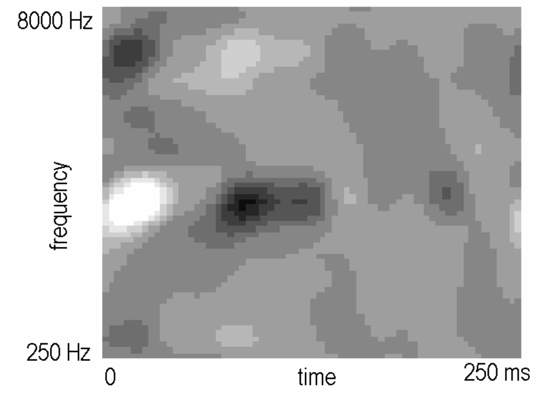

Figure 1 A cortical receptive field, observed in the auditory cortex of ferret (courtesy of David Klein, used with permission).

These relatively recent findings about the physiology of auditory cortex may have important implications in ASR. Since neurons in the auditory cortex respond best to certain kinds of acoustic signals (e.g. [31]), they seem to act as a kind of two-dimensional matched filter that detects the existence of a particular pattern (acoustic event) in the incoming signal. Then, just as a certain combination of formants indicate certain steady vowel sound; a certain combination of particular events could indicate certain dynamic acoustic event such as realistic dynamic phoneme of the language.

To elaborate this notion further, 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. A receptive filed responding to the instantaneous formant position is shown in the left part of Fig. 2.

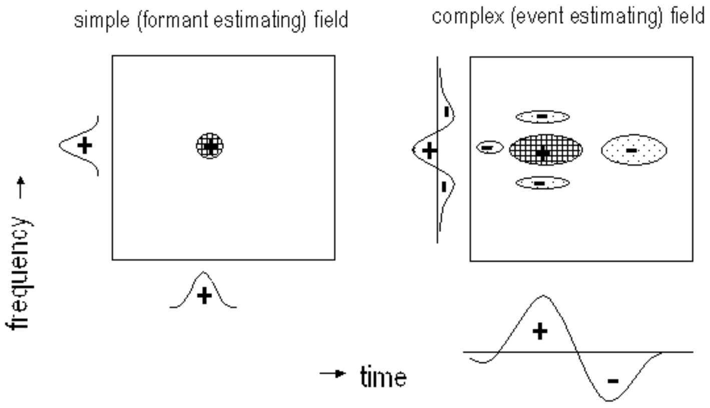
simple (formant estimating) field
complex (event estimating) field
frequency →
→ time
+
-

Figure 2 Schematic examples of simple and more complex receptive fields

Typical cortical receptive field is more complex than a single short excitatory region. 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 complex acoustic events.

Thus, to account for such complexity we need to replace formants by more complex spectro-temporal events as carriers of linguistic information in speech. Such broadly defined events are characterized by complex time-frequency patterns, involving times other than the current instant.

2. Data Guided Feature Extraction

The more knowledge we build into the feature extraction module, the less we need to train the subsequent stochastic recognizer. The question is where this knowledge should come from. Both techniques, the Mel cepstrum and the PLP have been designed by implementing some rudimentary textbook knowledge about auditory perception. As evidenced by the success of data-driven stochastic pattern classification and language modeling methods (see e.g. [29] for details), using incorrect prior knowledge may be worse than using no prior knowledge at all.

Why do we need to train the analysis module to derive features that will then be used in another trained stochastic system? Our knowledge about coding of information in speech signal is still incomplete. Enough speech data, labeled with respect to the targeted linguistic message (either by hand or by forced alignment procedures) is available and can provide the feature extraction module with speech-specific knowledge. Would it be also possible to derive the knowledge contained in the feature extraction module from the data?

It turns out that the answer is affirmative. For this purpose, we will be using two techniques, both aiming at providing an optimal set (in a sense of being the smallest) of features for a classification. The first one is the well known 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. [25, 5] for some examples of previous use of LDA in ASR), the second one is a nonlinear technique that uses a particular form of an artificial neural net called feed-forward Multi-Layer Perceptron (MLP) that, when properly trained, is capable of estimating posterior probabilities of classes of interest [6, 7].

2.1. Data-Derived Spectral Aspects of Information in Speech Signal

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 the spectral basis illustrated in Fig. 3. 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. [35, 36] show that the spectral resolution implied by spectral basis in Fig. 3 is very similar to spectral resolution of auditory-like Bark frequency scale. This finding supports earlier results [43] that derived auditory-like frequency warping by minimizing differences between speeches from different talkers.

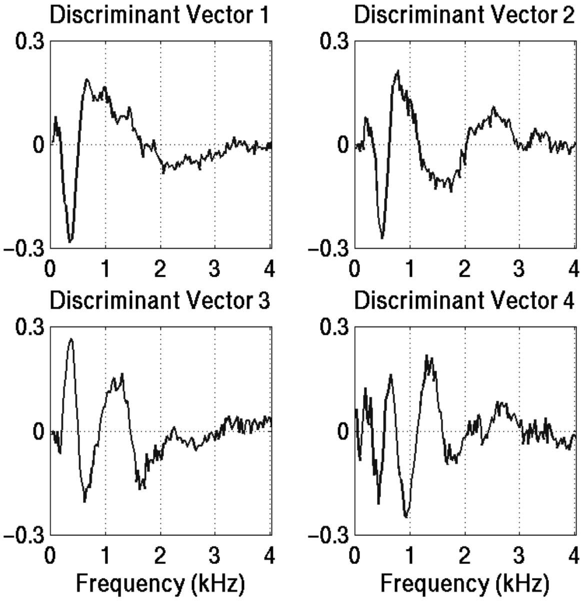

Figure 3 Spectral basis vector derived by data-driven LDA technique.

2.2. Data Derived Temporal Aspects of Information in Speech Signal.

RASTA processing filters the time trajectories of speech features to attenuate the features with rate-of-change that is not expected for speech. The initial ad hoc form of the RASTA filters [24] was optimized on a relatively small series of ASR experiments with noisy telephone digits. We later realized that it is possible 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 second) 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.

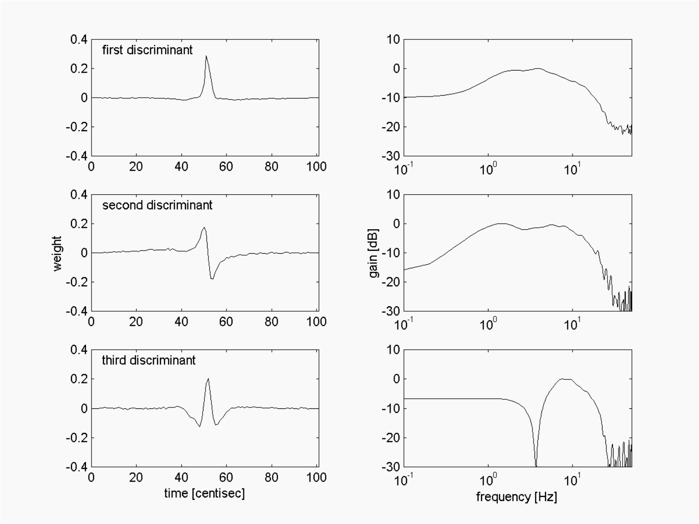

Figure 4 Impulse and frequency responses of the first three discriminant vectors from the LDA-derived discriminant matrix. The filters for the 5 Bark frequency channel are shown here. Filters for the other carrier frequencies studied (between 1 and 14 Bark) are very similar.

Frequency responses of the first three FIR filters derived from OGI Stories database are shown in Fig. 4. As seen in Fig. 5, filters for different frequency channels are similar.

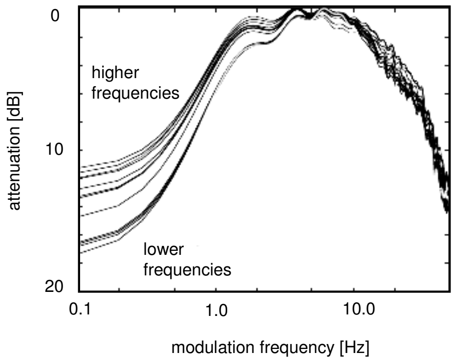

Figure 5. Frequency characteristics of the first discriminant derived from the American English portion of the OGI Stories database for all 15 critical bands

The frequency characteristic (shown at in the right part of the Figure) are generally consistent with RASTA [24], and delta, and double-delta feature of speech [14]. However, the 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. This gives a mild bandpass filter as shown in the upper right of the figure. The second discriminant vector computes the difference between weighted averages from left and right contexts of the current frame (the first derivative of the first discriminant vector). The third discriminant vector is 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 [37]. These three vectors correspond to RASTA filtering with subsequent computation of the first (delta) and the second (delta-delta) dynamic features [14].

2.3. LDA on much Larger Database

Experiments are run using 30 hours of speech obtained from the CTS (Conversational Telephone Speech) database. CTS database is a collection of narrowband speech data from many different previous databases (Switchboard, Fisher databases, etc.). Data are labelled into 40 phonetic classes and labels provided by SRI are automatically obtained using forced alignment. This amount of data is significantly larger than the amount of data previously used for this task, allowing robust estimation of Spectro-Temporal discriminants.

In this first set of experiments we apply LDA as described in section (2). Until now only first three discriminants have been studied; the use of large amount of data allows the robust estimation of higher discriminants. If data are split in sentences as it is usually done in speech recognition, discriminants exhibit artefacts at the end and at the beginning, otherwise if they are processed with full context they show significant non-zero values only in the centre. This suggests that the procedure of splitting the data in blocks may be detrimental for such temporal processing of data. Figure 6 show four discriminants obtained from the fifth critical band.

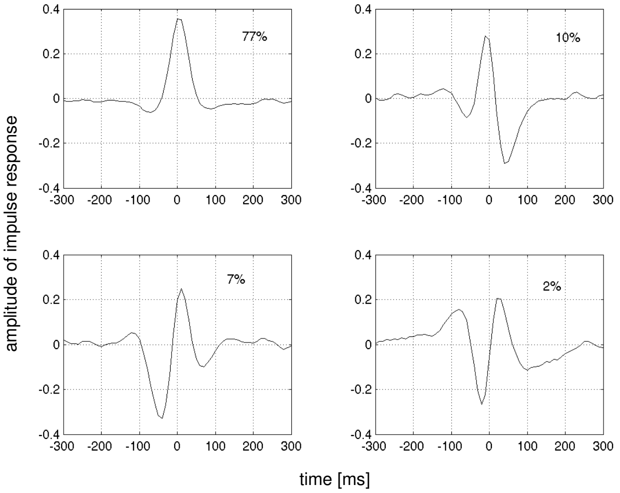

Fig 6a Impulse responses of the first four temporal discriminants derived by LDA on the Conversational Telephone Speech database

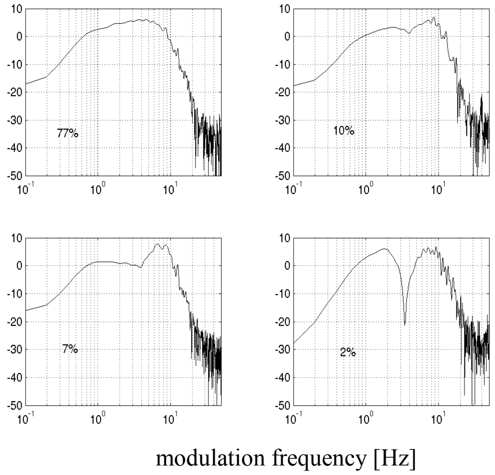

Figure 6b Impulse responses and frequency responses of the first four temporal discriminants derived by LDA on the Conversational Telephone Speech database

Discriminants at other frequencies are again very similar. Width of those filters suggests that information about phonemes is spread in time over an interval of around 500ms around the centre of the phoneme. First discriminant is qualitatively similar to RASTA filter while higher order discriminants describe more details of signal dynamics. In frequency domain they correspond to pass band filters that pass lower frequencies of the modulation spectrum. Width of temporal discriminants progressively increases, suggesting the use of different time resolutions.

In [23] spectral discriminants are derived using LDA, however PCA is used for smoothing the between-class and the within-class covariance matrixes, needed for deriving the LDA discriminant matrix [13]. According to the discussion of section 2 this is a suspect method that can significantly affect the result. We repeat the same experiment using 30 hours of speech and the previously discussed LDA technique for singular matrices. Hamming window shifted by 10ms step is used to obtain 129 points of 12th order LPC logarithmic power spectrum. Cross validation experiments select out of the possible 39 discriminants only 24 (that are enough for covering the discriminative space). These resulting discriminants are shown in the Fig. 7 below.

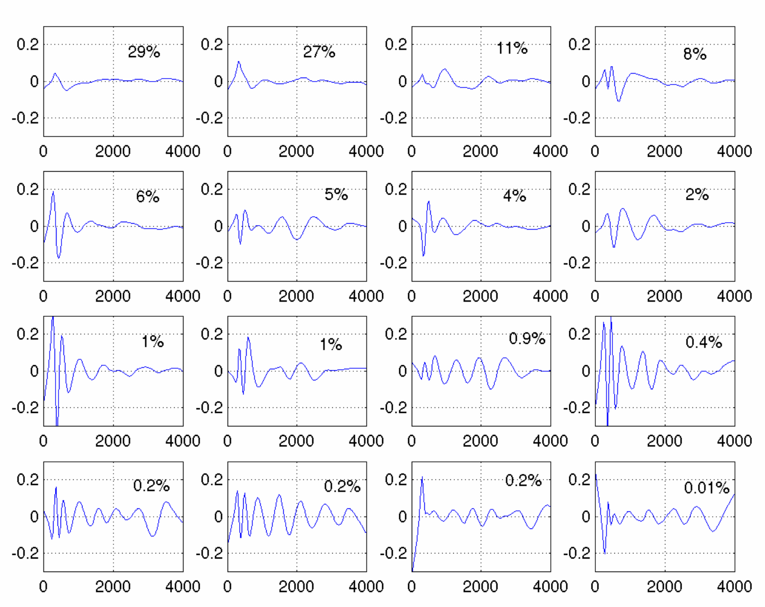

Figure 7 First sixteen spectral bases derived by LDA on the Conversational Telephone Speech database

Linear discriminants show a higher oscillation frequency at low frequency and progressively lower oscillation frequency in the higher part of the spectrogram, suggesting different frequency resolution at different parts of the spectrum. To further investigate this issue we performed sensitivity analysis as described in [55]. Sensitivity of a given bases is computed as the Euclidean distance between a gaussian shape centred at a given frequency and the same shape shifted by a certain value, projected on the bases. In other words if g(f) is a gaussian shape centred at frequency f and W is the LDA basis, the sensitivity S(f) is defined as S(f) = ||g(f) ·W −g(f +μ) ·W|| where μ is a shift. Figure 8 (left) plots sensitivity to a constant shift μ = 25Hz on a linear scale; in this case LDA basis are more sensitive at lower frequencies. Figure 8 (left) plots sensitivity to a constant shift μ = 0.8 Bark on a Bark scale; sensitivity is now constant, suggesting that LDA discriminants emulate the Bark scale with higher resolution at low frequencies and lower resolution at higher frequencies. The Bark scale was derived by perceptual experiments, LDA spectral basis are completely data-driven supporting [55].

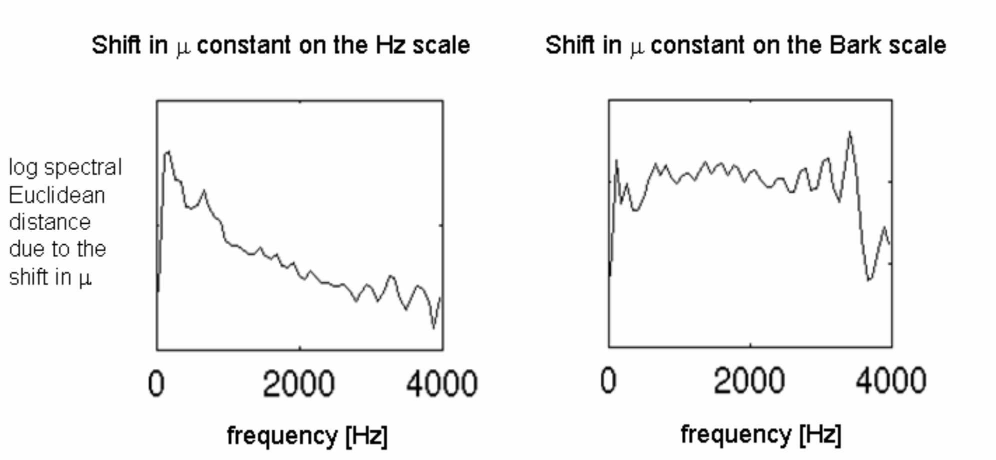

Figure 8 Sensitivity of the LDA-derived spectral projections to perturbances of a spectral component at different frequencies. Left, the perturbance step is constant on the linear spectral scale. Right, the perturbance step is constant on the nonlinear auditory-like Bark scale

2.4. Joint Spectro-Temporal Analysis

Analysis of spectrum and of temporal trajectories has been done in previous sections independently. We want to investigate now discrimination in joint spectro-temporal domain using discriminant analysis. A T ×F matrix where T is the temporal context and F is the number of frequency components is labeled according to the phoneme in its center. The matrix can be represented as a vector of size T × F and classical LDA can be applied. This approach was applied in [51] in critical band domain with a context of 101 frames on small amount of data. Conclusion was that the amount of data was not sufficient for robust estimation of discriminants. Better results were obtained if analysis was carried independently in time and in frequency and discriminants recombined. We are interested here in using the LPC power spectrum (dimension 129 points) in combination with a temporal context of 101 frames. If matrices are represented as vectors, S_w and S_b have a dimension of 13000 × 13000 which is unsuitable for computational reasons. A way to overcome this dimension problem is doing discriminant analysis directly in the matrix space represented as a tensor. In other words, operations on vectors are replaced by operations on tensors and final discriminant space is a tensorial space. If X is a matrix of dimension T × F, we seek the space transformation that reduces T × F into a space of dimensions $l_1 \times l_2$; this space is obtained by the tensor product of a subspace L of dimension T $\times l_1$ and R of dimension F × l_2. Projection of an element X in this space is given by the product $Y = L^TXR$ with final dimension $l_1 \times l_2$. In the tensorial space, the Frobenius norm can be used to derive within and across class matrices S_w and S_b defined as:

$$S_w = \sum_{i=1}^{k} \sum_{X \epsilon C_i} ||X - M_i||_F^2 \quad S_b = \sum_{i=1}^{k} n_i ||M_i - M||_F^2$$

where M_i is the class mean matrix and M is the global mean matrix. Using the Frobenius norm property trace(MM^T) = $||M|^2_F$ and applying the transform $Y = L^TXR$, expression reduces to:

$$\begin{aligned} \bar{S}_w &= trace(\sum_{i=1}^{k} \sum_{X \epsilon C_i} L^T (X - M_i) R R^T (X - M_i)^T L) \\ \bar{S}_b &= trace(\sum_{i=1}^{k} n_i L^T (M_i - M) R R^T (M_i - M)^T L). \end{aligned}$$

Optimal transforms L and R can be found iteratively fixing one of them, projecting tensor in their space and solving the generalized eigenvalue problem (see [53]). This result in eigen-decomposition of matrix 101×101 and 129×129 instead of 13000×13000. In other words rows are projected on matrix RRT and columns on matrix LTL that span the linear discriminant space of rows and columns of X respectively. Linear discriminants obtained using 2D-LDA have similar shape as outer product of discriminants obtained processing independently the temporal and spectral domain, suggesting that those domains can be processed independently. Figure 9 shows twelve 2D discriminants in the time frequency domain. Some of them show strong localization properties both in time and in spectral domains as if they were sensitive to a particular region in the plane; on the other hand we can notice as well discriminants with a sensitivity more spread over the frequency domain.

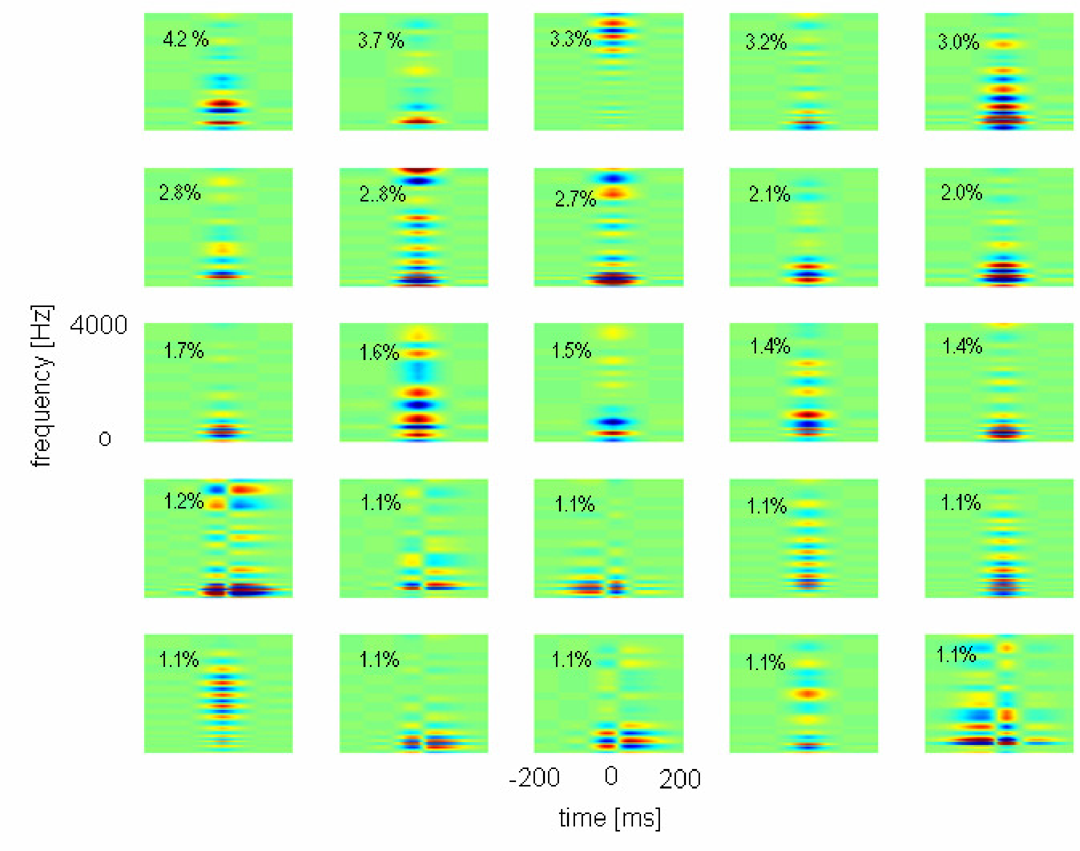

Figure 9 The first sixteen 2-D (spectro-temporal) bases derived by 2D-LDA on the Conversational Telephone Speech database

3. ASR Based on Complex Spectro-Temporal Patterns

Now, how do we use this 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 within state uncorrelated and normally distributed features every 10 ms or so. Further, the feature vectors should be low-dimensional so that the subsequent pattern classifier is also small and could be trained on a finite amount of training data. The smallest set of features for classification is posterior probabilities of the classes to be classified [13]. So we need a module capable examining relatively long spans of speech signal within various frequency bands 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.

3.1. Introduction to TRAP-TANDEM Technique

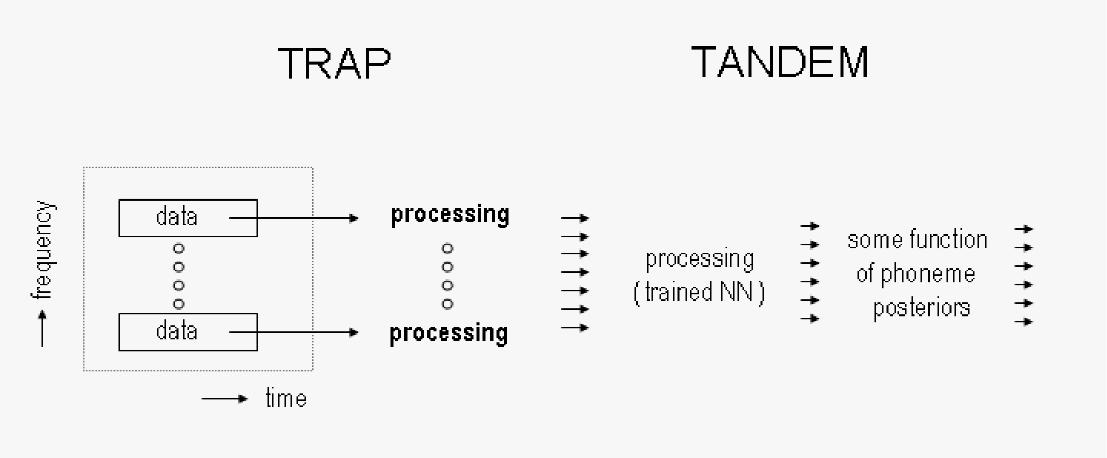

Figure 10: Trap-Tandem Feature Extraction

A step in this direction is the TRAP-TANDEM and related techniques [18, 20]. A schematic picture of the TRAP-TANDEM techniques is shown in Fig.10. The TRAP (standing for TempoRAl Pattern) 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 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. The name TANDEM reflects the fact that the classifier is used in tandem with the conventional HMM-based classifier. Both the TRAP and the TANDEM modules are trained on development data.

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 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 [26]) discriminative classifiers. Hierarchical classification schemes in TANDEM estimator were also investigated [42].

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 [4] and confirmed by Yang et al. [44]. 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 [33], hence 1 s.

Why should we abandon the short-term spectrum of speech? First, 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 [12] (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.

3.1.1 TRAP

The input to the TRAP module is formed from one or more time trajectories of critical-band energies. Some benefits are seen when more than one time trajectory is used as an input. At the moment, the time-frequency spectral density plane uses the front end module from PLP analysis [16]. It does the short-time spectral analysis of the speech signal with a subsequent Bark-like summation of the spectral components. However, a 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 [2]. Several ways of pre-processing in the TRAP module have been examined, two of these are described in some detail below

a) Pre-processing by trained speech-class posterior probability estimators

One possibility is to estimate frequency-local posterior probabilities of speech classes using trained non-linear frequency-local classifiers. TRAP estimators, using multi-layer perceptrons, 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 common American English phonemes clustered into 6 broad phonetic classes [1, 27] and a separate estimator is 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 [19]. Even when the nonlinear classifiers are used, it still seems to be advantageous to pre-process the input data prior to the frequency-localized classification. The PCA analysis of the data suggests that to preserve 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 [28, 15]. The most successful dimensionality reduction has been so far the cosine transform that typically allows for at least 50% reduction of the input data.

b) Hard-wired pre-processing

Another, more recently pursued possibility, is to pre-process different frequency-localized time-frequency patches by 2-D filters, operating on the critical-band modulation spectrum, and to feed the outputs of this bank of filters directly to the subsequent TANDEM module [13f]. The 2-D filters we are currently using are band-pass filters which impulse responses in the temporal domain represent the first and the second temporal derivatives of Gaussian functions

$$g_1[x] \propto -\frac{x}{\sigma^2}\ \exp(-\frac{x^2}{2\sigma^2}),$$

$$g_2[x] \propto (\frac{x^2}{\sigma^4} - \frac{1}{\sigma^2})\ \exp(-\frac{x^2}{2\sigma^2}),$$

where x is time with the step of 10 ms; standard deviation σ determines the effective width of the Gaussian. We have used 8 different values of σ, logarithmically spaced on the interval 8-130 ms. Filters with low σ values have finer temporal resolution, high σ filters cover wider temporal context and yield smoother trajectories. All temporal filters are zero-phase FIR fiters, i.e. they are centred around the frame being processed. Length of all filters is fixed at 101 frames, corresponding to roughly 1000 ms of signal, thus introducing a processing delay of 500 ms. First and second derivatives of Gaussian function have zero-mean by the definition. By using such impulse responses we gain an implicit mean normalization of the features within a temporal region proportional to the value of σ, which infers robustness to linear distortions [24]. Since such removal of the DC component from temporal trajectories of logarithmic spectral energies is one of the key properties of the RASTA processing [24], we call this particular processing the Multi-RASTA (MRASTA) processing. Impulse responses given by the upper equation are shown in the left part of Fig.11, the right parts shows impulse responses given by the lower equation. Respective frequency responses are illustrated in Fig. 12.

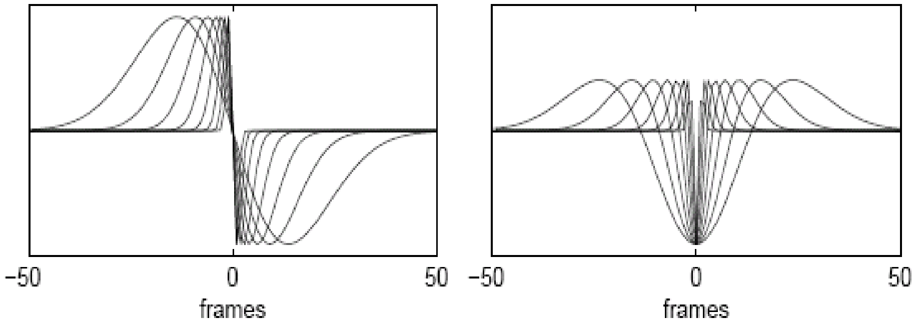

Figure 11.Normalized impulse responses of two sampled and truncated Gaussian derivatives for σ= 8 - 130 ms.

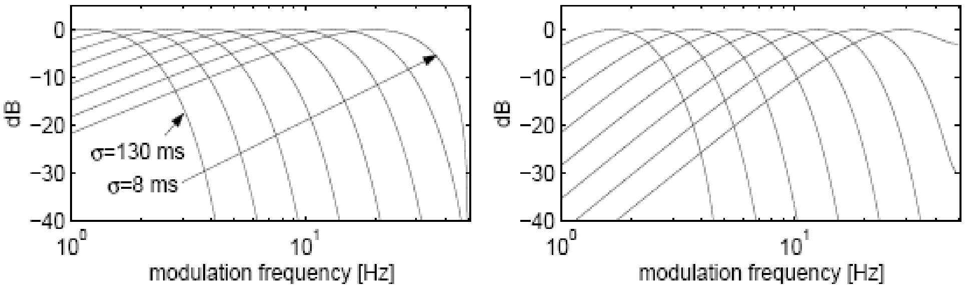

Figure 12. Normalized frequency responses of the two sampled and truncated Gaussian derivatives for σ = 8 - 130 ms.

Frequency derivatives at each critical band frequency were formed by subtracting filter outputs at the lower neighboring frequency from the output at the higher neighboring frequency at each time instant., thus in effect forming differentiating filter with its impulse in the frequency domain given by the vector {-1, 0, +1 }. In that way, 442 different 2-D time-frequency modulation domain filters were formed. An example of one of these filters is illustrated in Figure 13.

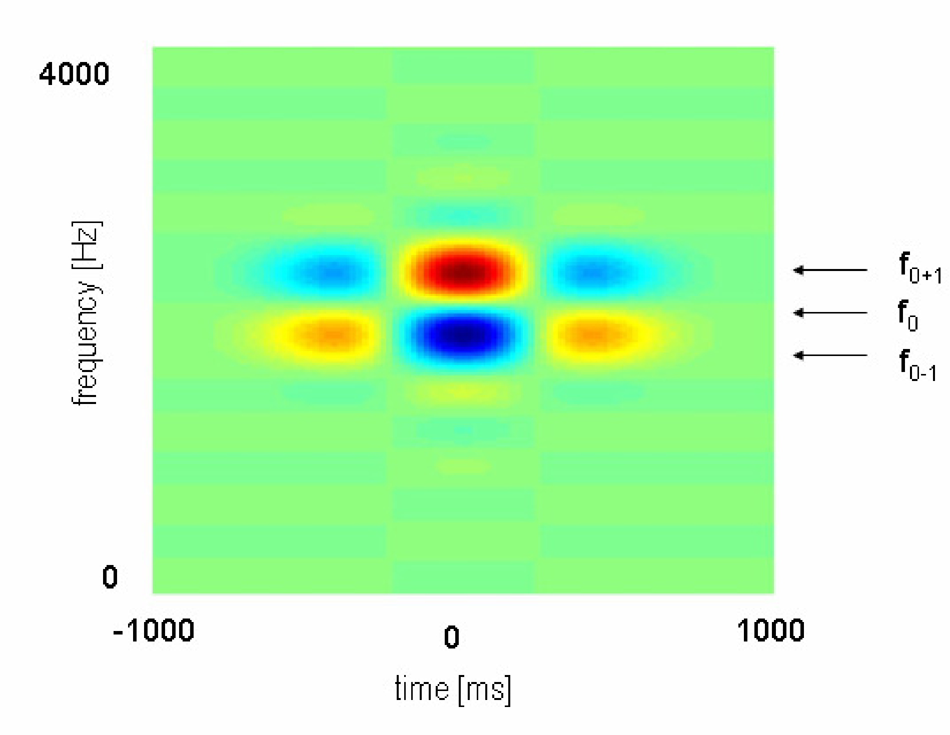

Figure 13. An example of an impulse response of one of2-D time-frequency modulation domain filters. Warm colors (yellow to red) indicate positive values, cold (blue) color indicate the negative ones, green indicates values close to zero. This particular filter represents the second derivative of the Gaussian function in the time domain and the first derivative in the frequency domain, centered at f_0 .

3.1.2 TANDEM

Techniques based on optimal rotation of feature space such as linear discriminant analysis (LDA) have been used in feature extraction in ASR for quite some time [25, 5]. A 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 [6,7].

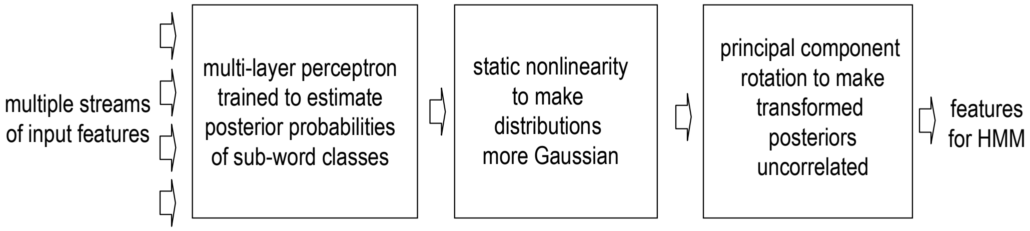
multiple streams of input features
multi-layer perceptron trained to estimate posterior probabilities of sub-word classes
static nonlinearity to make distributions more Gaussian
principal component rotation to make transformed posteriors uncorrelated
features for HMM

Figure 14 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, in TRAP-TANDEM 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 [3, 34].

4. Autoregressive modeling of temporal trajectories of spectral energies in frequency sub-bands

The speech signal is not stationary but carries information it its dynamics. To enable the use of processing techniques that assume signal stationarity, short segments of the signal (10-30 ms) are used to derive short-term features for pattern classification in automatic speech recognition (ASR). The signal dynamics are then represented by a sequence of the short-term feature vectors with each vector representing a sample from the actual underlying dynamic process, in a manner similar to the way motion in movies is represented by a sequence of static shots. The issues of windowing, time-frequency resolution compromises, proper sampling of the short term representation, emulating the unequal frequency resolution of hearing, etc., are typically addressed in an ad hoc manner. To parameterize short-term spectral envelopes, a rich inventory of techniques has evolved.

The temporal resolution of such frame-based representation is the same at all frequencies and is given by the applied analysis window (typically around 25 ms) which acts as a low-pass filter on the temporal trajectories.

Unlike the conventional feature extraction techniques, our features are based on temporal evolution of spectral energy in frequency sub-bands. We are therefore interested in finding alternatives to the conventional frame-based analysis where we could directly model temporal aspects of the time-frequency plane and to control temporal properties of the model. To begin our quest, we recall that an alternative way of deriving the short-term speech representation (applied e.g. in the original Spectrograph) could be using the rectified output from a bank of band-pass filters. Spectral resolution could be then controlled by band-pass filter design, and the temporal resolution could be different at different frequencies depending on the lengths of impulse responses of the individual filters.

There is a third, perhaps less obvious way of deriving the short term spectral representation. Just as a squared Hilbert envelope (squared magnitude of the analytic signal) represents instantaneous energy in a signal, the squared Hilbert envelopes of the sub-band signals are a measure of the instantaneous energy in the corresponding sub-bands. To get the Hilbert envelope would normally involve the use of either the Hilbert operator in the time domain (whose infinite impulse response presents some practical issues) or the double use of the Fourier transform with modifications to the intermediate spectrum [73]. An interesting and practical alternative is to get the all pole approximation of the Hilbert envelope by computing a linear predictor on the cosine transform of the signal. Such Frequency Domain Linear Prediction (FDLP) is the frequency domain dual of the well-known time-domain linear prediction (TDLP). In the same way TDLP fits an all-pole model to the power spectrum of a signal, FDLP fits an all-pole model to the squared Hilbert envelope. Since the cosine transform represents the Fourier transform of the even-symmetrized time signal, the “spectrum” of the resulting predictor gives an approximation to the Hilbert envelope of the signal (in the same way as the spectrum of the predictor derived in the time domain is an approximation of the power spectrum of the signal).
To get an all-pole approximation of the Hilbert envelope for a specific sub-band, the prediction needs to be derived only from the appropriate part of the cosine-transformed signal. The parametric all-pole description of the temporal trajectory offers control over the degree of smoothing of the Hilbert envelope.

The easily-computable “cepstrum” of the time-domain all-pole model represents in this case the spectrum of the logarithmically-compressed temporal envelope and is related to the cosine transform of the original TRAP which has been found useful in ASR [74]. The duality between the power spectrum and the squared Hilbert envelope is essential to the understanding of FDLP.

Figure 14 illustrates these two dual forms of linear prediction. On the upper left pane we display 50 ms of speech that we want to model using the two dual forms of linear prediction. Conventional linear prediction (TDLP in our terminology) approximates the power spectrum of the signal, as shown in the middle panel on the right (frequency) side of the figure, which is the TDLP of the top-left (time) signal. The full Fourier power spectrum to which this is an approximation is plotted directly below, in the bottom-right pane. FDLP on the other hand operates on the DCT of the signal (top right pane) and results in an LP model describing the *temporal* envelope, shown in the middle left (time) panel. Each column provides three alternative representations of each domain. Whereas TDLP exploits the spectral structure of the signal to construct an efficient predictor of the temporal signal, FDLP exploits the temporal structure of the signal to predict spectral values.

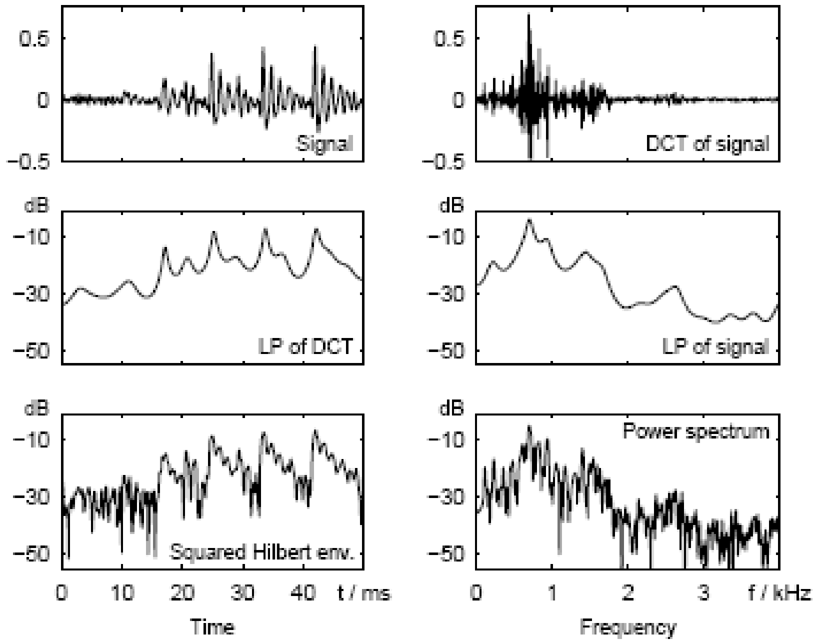

Figure 14: *The two dual forms of linear prediction.* On the left column (time) we plot 50 ms of a speech signal, the FDLP all-pole fit and corresponding squared Hilbert envelope. On the right column (frequency) we display the DCT of the same signal, the conventional time-domain LP all-pole fit and the corresponding power spectrum. Both models use 28 poles.

4. 1. Mathematical description of envelope estimation

This section provides mathematical description of the steps involved in the estimation of the temporal envelope in more detail. The concept of FDLP was to our knowledge first introduced by Herre [72] as a method for efficient coding of transients in transform coders. Kumaresan has independently discovered and extensively worked on FDLP, a method which he calls linear prediction in the spectral domain or LPSD [75].

In our work, we apply FDLP to approximate relatively long temporal envelopes of the sub-band signal. However, to simplify the notation in this section, we present the full-band version of the technique. The sub-band based technique is identical but applied only to the appropriate parts of the DCT transformed signal.

Let us define the input discrete time-domain sequence as *s(n)*for time samples *n =1,...,N,* where *N* denotes the segment length. Its Fourier power spectrum $P(\omega_k)$ (sampled at discrete frequencies $\omega_k = 2\pi k/N$, k =1,...,N) is given as

$$P(\omega_k) = |S(e^{j\omega_k})|^2,$$

where $S(e^{j\omega k}) = Z\{s(n)\}|_{z=e} j\omega_k$. $Z\{ .\}$ stands for the z-transformation. Later, let us use the notation $F\{.\}$ for Discrete Fourier Transform (DFT) which is equivalent to *z*-transform with $z = e^{j\omega k}$. It has been shown, e.g., in [76], that classical Temporal-Domain Linear Prediction (TDLP) fits the discrete power spectrum of an all-pole model $\hat{P}(\omega_k)$ to $P(\omega_k)$of the input signal.

Unlike TDLP, where the time-domain sequence *s(n)* is modelled by linear prediction, FDLP applies linear prediction on a frequency-domain sequence. In our case, *s(n)* is first DCT transformed. It can also be viewed as the symmetrical extension of *s(n)* so that a new time-domain sequence *q(m)* is obtained *(m =1,..., 2N)* and then DFT projected. We obtain the real-valued sequence $Q(\omega_k) = F\{q(m)\}$. We then estimate the frequency-domain prediction error $E(\omega_k)$as a linear combination of $Q(\omega_k)$consisting of p real prediction coefficients b_i

$$E(\omega_k) = Q(\omega_k) - \sum_{i=1}^{p} b_i Q(\omega_k - i).$$

The b_i are found so that the squared prediction error is minimized [76]. As noted above, in the case of TDLP, minimizing the total error is equivalent to the minimization of the integrated ratio of the signal spectrum $P(\omega_k)$to its model approximation $\hat{P}(\omega_k)$

$$E_{TDLP} \approx \frac{1}{N} \sum_{k=1}^{N} \frac{P(\omega_k)}{\hat{P}(\omega_k)}.$$

In the case of FDLP, we can interpret $Q(\omega_k)$ as a discrete, real, causal, stable sequence (consisting of frequency samples). Its discrete power spectrum will be estimated through the concept of discrete Hilbert transform relationships [77]. $Q(\omega_k)$can be expressed as the sum of $Q^e(\omega_k)$and $Q^o(\omega_k)$, denoting an even sequence and an odd sequence, respectively; thus $Q(\omega_k)= Q^e(\omega_k)+ Q^o(\omega_k)$. Its Fourier transform

$$\phi(m) = F\{Q(\omega_k)\} = \phi^R(m) + j\phi^I(m),$$

where R and I stand for real and imaginary parts of φ(m), respectively. It has been shown (e.g., [77]) that $\varphi^R(m)= F\{Q^e(\omega_k)\}$ and $\varphi^I(m)= F\{Q^o(\omega_k)\}$. By taking the Fourier transform of $Q^e(\omega_k)$, the original sequence $q(m)$ is obtained

$$F\{Q^e(\omega_k)\} = \phi^R(m) = const\ q(m).$$

The relations between $F\{Q^e(\omega_k)\}$and $F\{Q^o(\omega_k)\}$, called the Kramers-Kronig relations, are given by the discrete Hilbert transform (partial derivatives of real and imaginary parts of analytic function [78]), thus

$$\phi(m) = \phi^R(m) + j\phi^I(m) = const\big(q(m) + jH\{q(m)\}\big),$$

where $H\{ .\}$ stands for Hilbert transformation. Power root $|\varphi(m)|^2$ is called the squared Hilbert envelope. Prediction error is proportional to the integrated ratio of $|\varphi(m)|^2$ and its FDLP approximation $A(m)^2$

$$E_{FDLP} \approx \frac{1}{2N} \sum_{m=1}^{2N} \frac{|\phi(m)|^2}{A^2(m)}.$$

This equation can be interpreted in such a way that the FDLP all-pole model fits squared Hilbert envelope of the symmetrically extended time-domain sequence $s(n)$. FDLP models the time-domain envelope in the same way as TDLP models the spectral envelope. Therefore, the same properties appear, such as accurate modeling of peaks rather than dips. Further, the squared Hilbert envelope $|\varphi(m)|^2$ is available and can be modified. Thus, e.g., compressing $|\varphi(m)|^2$by a root function $[.]^r$ turns it into

$$E_{FDLP} \approx \frac{1}{2N} \sum_{m=1}^{2N} \frac{|\phi(m)|^{\frac{2}{r}}}{A^{\frac{2}{r}}(m)}.$$

In our experiments, the DCT input sequence is weighted by a set of Gaussian windows of variable temporal resolution, spaced following the Bark scale, as described in [70]. Gaussian windows span the whole DCT sequence. Therefore, we can individually exploit FDLP in each critically band-sized sub-band.

4.2. Linear predictive temporal patterns (LP-TRAP)

In our experiments we extend the FDLP model to speech segments up to 1 sec long. We seek here to summarize the temporal dynamics rather than capture every single nuance of the temporal envelope. Taking the DCT of a 1 sec speech segment at 8 kHz sampling rate generates 8000 frequency domain samples. Instead of fitting one predictor on the whole frequency series as we do in figure 1, we first apply 15 Bark-spaced overlapping Gaussian windows. We then apply FDLP separately on each of the 15 bands. Each predictor then approximates the squared Hilbert envelope of the corresponding sub-band. This is the "sub-band FDLP" introduced in [2] but here we extend the time window to even longer speech segments and use overlapping windows. We compute the auditory spectrogram over the 1 sec windows by stacking the individual temporal trajectories (rather than by stacking the individual frequency vectors as done in the conventional short-term spectral analysis). This is demonstrated in figure 15.

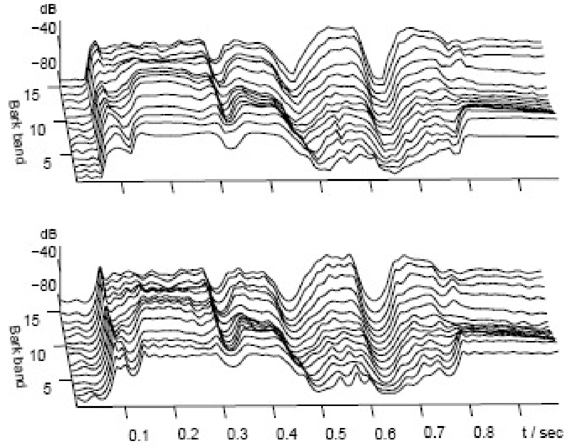

Figure 15: *Auditory spectrogram versus all-pole trajectories.* The first panel displays the short-time auditory spectrogram whereas the second panel shows the FDLP-approximated Hilbert envelopes using 80 poles per band. below it is plotted the corresponding squared Hilbert envelope that is being estimated.

The top panel shows the auditory spectrogram obtained by short-term Fourier transform analysis and Bark scale energy binning to 15 critical bands. In the second panel we fit fifteen 80-pole FDLP models, one for each Bark band, and display the 15 estimates of the squared Hilbert envelopes.

4.3. Spectral transform linear prediction (STLP)

Spectral transform linear prediction was introduced as method to adjust the relative fit of the conventional (TDLP) predictor to the peaks and dips of the speech spectrum [69]. By raising the power spectrum to an arbitrary power, the *compression factor*, one can adjust the peak-hugging property of linear prediction. STLP is an integral part of the well-known perceptual linear prediction (PLP) technique; the cube-root compression of the power spectrum in PLP prior to prediction is an instance of STLP with compression factor 1/3. We borrow this idea and we apply it on the Hilbert envelopes of the Bark sub-bands instead of the power spectra. In some sense this method could now be dubbed temporal transform linear prediction (TTLP). In figure 16 we demonstrate the effect of the compression factor. We take the sixth Bark band from figure 2 and this time we keep the number of poles fixed to 50. On the top pane we display the logarithm of corresponding squared Hilbert envelope. On the second pane we plot the FDLP sub-band envelope with compression factor 1.0 which amounts to no compression. The all-pole model fits the peaks much better than the dips. A moderate compression of 0.1 still gives a better model of the peaks but of a greatly compressed Hilbert envelope. This time the dips are much better modeled. Lastly for compression of -0.5 the all-pole model fits a compressed version of the *inverse* Hilbert envelope. The dips are now accurately modeled. Spectral expansion using compression factors greater than 1 is also possible but it may result in ill-conditioned solutions due to the extreme sharpness of the peaks; we do not consider spectral expansions here.

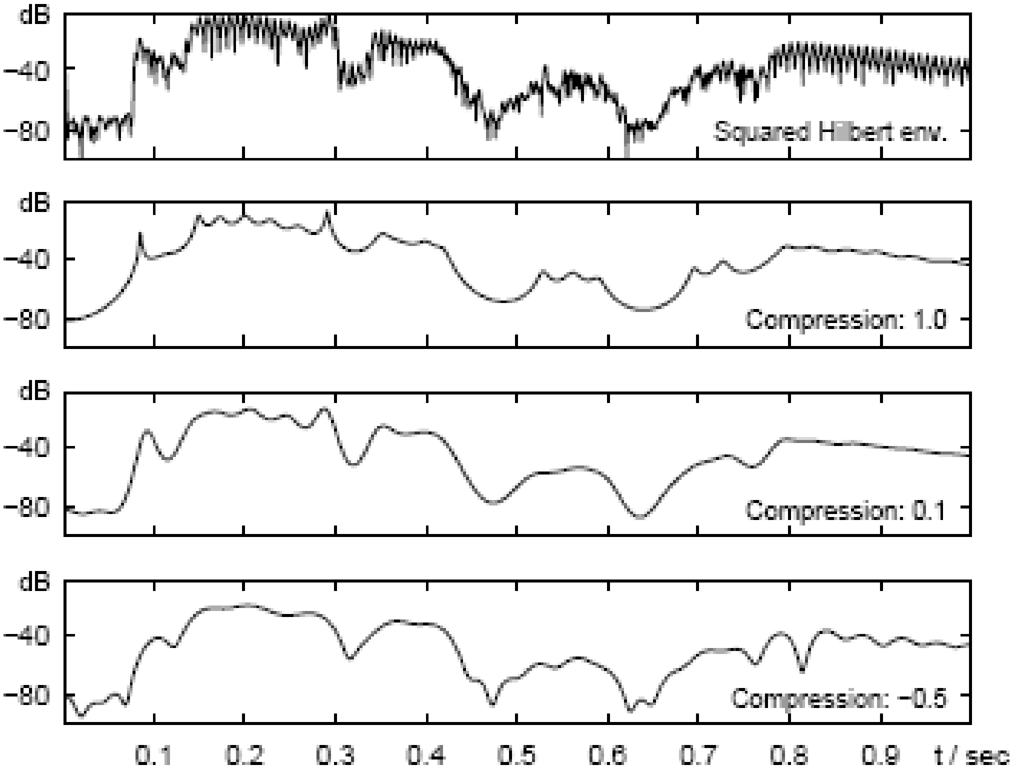

Figure 16: *The effect of the compression factor.* The top pane shows the squared Hilbert envelope of the sixth Bark band of figure 2. In the second pane FDLP using no compression fits the peaks. Using moderate compression in the third pane FDLP achieves a better fit to dips in the envelope. The fourth model fits a compressed version of the inverse spectrum, thereby fitting the dips in preference to the peaks. The number of poles is 50 in all three cases.

For comparison, we show in Fig. 17 2-D spectro-temporal representations of the same utterance derived by a conventional frame-based feature extraction technique (PLP analysis) and by the proposed FDLP. A typical spectrogram is constructed by appending individual short-term spectral vectors alongside each other. A similar representation can be constructed by vertical stacking of the temporal vectors approximating the individual sub-band Hilbert envelopes. The top panel in Fig. 4 shows the result of PLP smoothing, with each 15-point vertical spectral slice now smooth and continuous as a result of being fit with an LP model. The bottom panel is based on a series 24-pole FDLP models, one for each Bark band, to give estimates of the 15 subband squared Hilbert envelopes. As with PLP, cube-root compression is applied here to the sub-band Hilbert envelope prior to computing the all-pole model of the temporal trajectory. The similarity of all these patterns is obvious, but there are also some important differences: As discussed above, the FDLP technique is capable of extracting finer temporal details that are smother out in the PLP results due to its stationarity assumption with the 20 ms analysis interval applied.

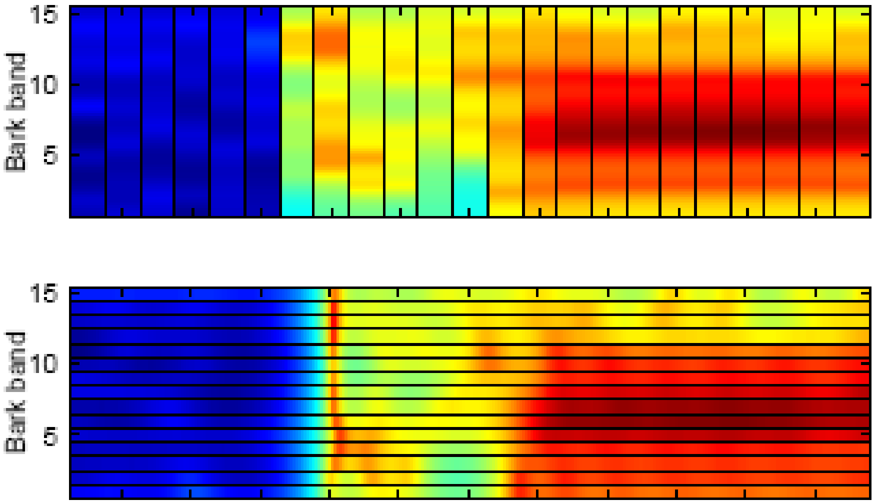

Fig.17 Spectro-temporal representation (spectrogram) from conventional frame-based analysis (PLP with 20 ms Hamming window and 10 ms analysis step) and from FDLP

5. Speech/Non-speech detection from long-temporal spans

The speech processing technique utilizing limited frequency ranges and longer temporal spans has been successfully employed in one of the crucial speech signal processing task – speech/non-speech detection (SND). By providing robust and accurate speech detection, ASR performances can be highly improved, especially when dealing with challenging input data, such as e.g. meeting recordings. Inaccurate boundaries are an important cause of errors in automatic speech recognition systems, and a pre-processing stage that segments the signal into periods of speech and non-speech is invaluable in improving the recognition accuracy. An evaluation of an isolated-word recognizer has shown that more than half of the recognition errors are due to inaccurate word boundaries [57].

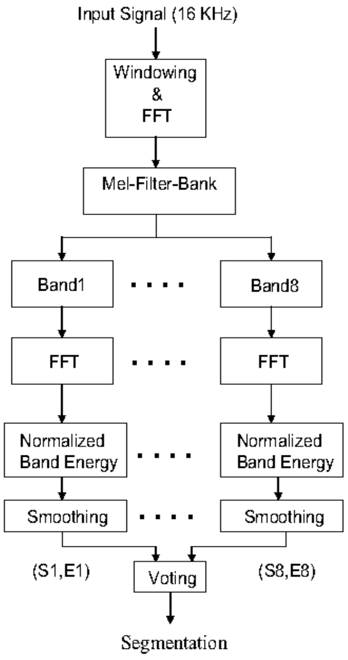

Fig. 18 Technique for speech/nonspeech discrimination based on modulation spectra derived from long temporal spans of band-limited spectral energies

One of the issues in the design of a SND system is the selection of an appropriate feature set that captures the temporal and spectral structure of the signals. Scheirer and Slaney investigated features for speech/music discrimination that are closely related to the nature of human speech [59]. The proposed features, including, spectral centroid, spectral flux, zero-crossing rate, 4 Hz modulation energy (related to the syllable rate of speech), and the percentage of low-energy frames, have been explored in the task of discriminating speech from various types of music. In [58], entropy and dynamism features based on posterior probabilities of speech phonetic classes (as obtained at the output of an HMM/ANN large vocabulary continuous ASR system) are used to form an observation vector sequence, which is used in a HMM classification framework.

The existing state-of-the-art methods can be split into two broad categories: *threshold detection process*, and *pattern-recognition process*. However, the both categories of SND systems are limited by two common drawbacks. On one hand, threshold based detection techniques fail under low SNR conditions, and on the other hand, pattern-matching techniques require large training data to train the models and need a prior knowledge of the noise.

The proposed approach is based on long-term modulations, examining the slow temporal evolution of the speech energy with time-windows in the range of 200 - 800 ms, contrary to the conventional short-term modulations (frequently used in ASR) studied with time-windows up to 10 - 30 ms which capture rapid changes of the speech signals. The relative prominence of slow temporal modulations is different at various frequencies, similar to perceptual ability of human auditory system. Particularly, most of the useful linguistic information is in the modulation frequency components from the range between 2 and 16 Hz, with dominant component at around 4 Hz [60-62]. In [61], it has been shown that for some realistic environments, the use of components from the range below 2 Hz or above 16 Hz can degrade the recognition accuracy. The proposed algorithm is based on this particular characteristic of speech, which is used to classify speech and non-speech signals in order to characterize each acoustic event.

6. Experiments

In first set of experiments we have compared LPC power spectrum that was projected on spectral and temporal basis. Results are compared with LPC power spectrum projected on DCT basis and with PLP cepstral coefficients. For experiments we used a database that is different from the one used for deriving linear discriminants assuming that those findings are universal properties of speech and not task dependent. Recognition results are run on the OGI-digits database.

Table 1 shows results obtained using the following set of 13 features: (a) LPC power spectrum projected on 13 DCT basis (b) PLP (c) LPC power spectrum projected on 13 spectral linear discriminants (d) LPC power spectrum projected on 13 spectral linear discriminants and filtered with one temporal discriminant. If LDA basis are used instead of DCT basis, an improvement of 4% (relative) is obtained. DCT basis has a uniform spectral sensitivity while LDA has a higher sensitivity at lower frequencies (emulating somehow the bark scale) where the most important information for recognition is contained. Spectral basis designed from data yield similar performance as PLP features designed according to auditory principles [10]. If spectral features are filtered with first temporal discriminant a considerable improvement of 35% (relative) w.r.t. the LPC baseline and 32% (relative) over PLP is obtained indicating the effectiveness of larger temporal context, imposed by the temporal filtering.

In table 2 we compare results for 39 features, i.e. LPCC features plus delta (an estimate of the first differential of temporal trajectory of the coefficient) and double delta (an estimate of the second differential) with power spectrum projected on 13 spectral basis and 3 temporal basis. In this case LDA spectral and temporal discriminants outperform LPCC plus delta and double delta by 11% (relative) while only very small improvement w.r.t. PLP and derivatives is found. Again data guided features yield equivalent results as currently often used PLP static and dynamic features.

Table 3 shows results for the TANDEM-based features. This time, the PLP-based TANDEM uses 9 frame of PLP+delta+ddelta features (the multi-frame input is beneficial in conjunction with the neural net-based TANDEM). The TRAP-TANDEM reaches the same performance.
However, the TRAP-TANDEM features have been so far found most useful in combination with the conventional spectrum-based (PLP, Mel Cepstrum,..) features. Thus, e.g. they were successfully applied in DARPA EARS program, where they brought about 10% relative improvement in error rate [39]. The MRASTA-TANDEM is significantly better (more than 20% relative improvement in the word error rate) even on its own. Combining it with more conventional PLP features yields yet further improvements.

Table 1: Word error rates for different sets of 13 features on OGI-digits

13 LPCC	13 PLP	13 spec.	13 spec. × 1 temp
14.1 %	13.5 %	13.5 %	9.1 %

Table 2: Word error rates for different sets of 39 features on OGI-digits

39 LPCC + differentials	39 PLP + differentials	13 spect. × 3 temp
6.0 %	5.4 %	5.3 %

Table 3: Error rates for TRAP-TANDEM and MRASTA-TANDEM features on OGI-digits

PLP_TANDEM	TRAP-TANDEM	MRASTA-TANDEM
4.5 %	4.5 %	3.5 %

Finally, we have also started to use FDLP features as input to TANDEM system. In this case, we have not used dynamic features (the first and the second differentials of temporal trajectories) do the baseline performance is lower (5.9 % word error rate as opposed to 4.5 % in the case when the differentials are used).

Optimal parameters of the FDLP model have been found earlier [81] as compression = 0.1, number of poles = 50, TRAP-length = 500ms, Gaussian window and applied here

Table 4: Experiments with FDLP-based feature extraction

features	Word error rate [%]
Baseline PLP features	5.9
Frequency response of FDLP as features	5.4
Cepstral coefficients of FDLP as features	5.2

Evaluation of speech/nonspeech detecting system was carried out on meeting recordings. We used a standard database containing the data recorded in an instrumented meeting room comprising of a microphone array, and headset and lapel microphones. All the microphones are of high quality electret type. The sensor configuration is similar to the system presented in [67]. The proposed technique was tested in conditions where the signal-to-noise ratio (SNR) varies considerably, as in the cases of close-talking headset, lapel, distant microphone array output, and distant microphone.

For a given 16 kHz sampled signal the Fast Fourier Transform (FFT) is computed over N points and the segment is shifted by n ms, resulting in N/2 dimensional FFT vector. The Mel-scale transformation is applied to the FFT vector. The filters used in Mel-frequency analysis are generally triangular in shape, and are equally spaced along the Mel-scale. The output is a Mel-scaled vector consisting of K bands. The computations are made over the entire incoming signal, resulting in a sequence of energy magnitudes for each band sampled at 1/n Hz. In each band, the modulations of the signal are analyzed by computing FFT over the P points and the segment is shifted by p ms. The result is a sequence of P/2 dimensional modulation vectors. The energies for the frequencies between the 2 - 16 Hz represent important components for the speech signal.

In case of SND, the experiments are conducted on a subset of the Multi-Channel Wall Street Journal Audio-Visual (MC-WSJ-AV) corpus. The specification and structure of the full corpus are detailed in [66]. A part of *Single speaker stationary* data, in which the speaker reads out sentences from different positions within the meeting room is used. Most of the data comprised non-native English speakers with different speaking styles and accents. The data is divided into development (DEV) and evaluation (EVAL) sets with no common speakers in both the sets.

To compare the efficiency of the proposed algorithm, short-term energy, short-term energy and zero-crossing based segmentation techniques [63], and a recently proposed MLP based system \cite{dines} are evaluated. This system relies on a MLP classifier, trained from several meeting room corpora to identify speech/non-speech segments. The training is performed with a corpus comprising headset recordings, which include approximately 112 hours of speech over 150 meetings.

To evaluate the proposed method, the parameters, mentioned above, are set as follows: N = 512, n = 10, K = 8, P = 100, p=10. The average energy in the 2-16 Hz for speech is approximately around 40 % of the total energy.

For the speech recognition experiments to evaluate the performance of the above mentioned techniques, a full HTK based recognition system [64], trained on the original Wall Street Journal database (WSJCAM0) was used. 52-element feature vectors were used, comprising of 13 MFCCs (including the 0th cepstral coefficient) with their first, second, and third order derivatives. Cepstral mean normalization is performed on all the channels. To reduce the channel mismatch between the training and test conditions, the baseline HMM models are adapted using a maximum likelihood linear regression (MLLR).

Speech recognition experiments on different channels including headset, lapel, distant microphone and the output of the beamformer as obtained from [67] are performed to evaluate the performance of the various techniques. The obtained results are shown in Table 4, where M, E, E+ZC, MLP, MS represent manual, energy, energy + zero-crossing, multi layer perceptron, and modulation spectrum based segmentations, respectively. The values in the first column represent baseline word error rate, which are obtained from the manual segmentation of the speech data. All the other values are compared with respect to these values. From the Table 4, it is clear that energy based approach (method E) performs poorly for all the channels. Adding the zero-crossing feature to energy (method E+ZC) helps in reducing the WER by about 50 % in all the cases. The MLP based approach performs close to manual segmentation as it is trained on headset data of the large corpus [65]. However, the performance decreases as the same MLP (headset trained) is used for lapel, distance microphone, and microphone array output for obvious reasons. From the table, it is also clear that the proposed modulation spectrum based approach is accurate and close to manual segmentation for all the channels.

Table 4: Word error rates for speech recognition system exploiting SND technique. The values in the first column represent baseline WER, obtained from manual segmentation, and other values are compared with respect to these values.

SIGNAL	M	E	E+ZC	MLP	MLP
Headset microphone	21.3	12.8	6.1	0.6	0.8
Lapel microphone	27.9	11.4	4.8	1.8	0.6
Distant Microphone	38.6	8.0	5.3	7.2	0.4
Beamformer Output	26.8	6.5	4.1	2.4	0.3

7. Discussion and conclusions

Short-term spectrum of speech and its spectral envelopes have been basis of features for ASR since its beginning. Gradually, auditory-like modifications of is frequency resolution [38] and of its amplitude axis [16], together with attempts for describing temporal dynamics of spectral envelopes [14] have emerged and been accepted by the engineering community. The current article shows that these modifications are supported by the character of speech signal. Namely, the nonlinear (critical-band like) spectral resolution as well as the description of spectral dynamics by local temporal derivatives emerges from the phoneme-based linear discriminant analysis of speech spectro-temporal plane. Moreover, these analyses also show the dominance of low (1-12 Hz) modulation frequencies in coding the phoneme-related information in speech with the subsequent need for relatively long (up to 1 second) segments of the speech signal when extracting this information from the signal. Further speculative reasoning then leads to abandoning the spectral envelope altogether, replacing it by frequency-specific posterior probabilities of speech-related events.

Further, the LDA analysis of spectro-temporal domain was done on quite different and much larger database of conversational speech . A tensorial LDA is proposed for processing long time-frequency slices and a revisited LDA is used for dealing with singular covariance matrices. Temporal basis have similar magnitude frequency characteristic as RASTA filters but differ in phase, spectral bases have similar frequency sensitivity as the Bark scale of hearing and obtained 2D filters show localization properties both in time and frequency. Those conclusions are qualitatively consistent with what was presented earlier in literature [23, 49, 55] on smaller databases. We found a large improvement in the use of data driven front-end when only 13 features are used. In this case the most important gain in performances is obtained when time trajectories are filtered with first temporal discriminant. On the other side only small improvements are obtained when dynamic features are added. The fact that results were carried over different databases supports the universal (speech specific and not task specific) nature of our findings.

That all then leads to a new data-driven feature extraction technique called TRAP-TANDEM which derives ASR features related to posterior probabilities of context-independent phoneme classes. 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 coming from beliefs and convictions of the designer but is 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 this way, the subsequent pattern classification module needs to learn only the task-specific knowledge, possibly reducing the need for the re-learning the same knowledge again and again each time the task changes. Derived features do not represent the 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 come from relatively short segments of speech representing a short part of the underlying sub-word class (phoneme). Instead the time span employed covers the typical coarticulation span of the phoneme. In that way, each feature vector carries 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 can be smaller and the burden on the subsequent HMM classifier is reduced. Unlike the conventional feature extraction approaches, it is consistent with the current knowledge of higher cognitive levels of mammalian auditory perception. The TANDEM-TRAP technique is still evolving and in order to get the most out of it, it may require some evolution in the existing ASR approach. However, it is already beneficial for the existing mainstream HMM-based ASR

A technique based on modulation spectrum from long temporal spans was employed in Speech/Non-speech detection task. This technique has been compared to manual segmentation, short-term energy, short-term energy and zero-crossing based segmentation techniques, and a recently proposed MLP classifier trained system. The speech recognition based evaluations are performed on real data in a meeting room for stationary speaker for all the methods and varying signal-to-noise ratios i.e. headset, lapel, distant microphone, and beamformer output. The results illustrate that the proposed simple technique is accurate, robust, close to real-time, and can be applied for SND tasks for any mode of speech acquisition and unforeseen conditions. Our study also raised a number of issues, including the approaches for decision without using the evaluation data (presently mean of the smoothed normalized energy is used), and the number of parameters involved in the method to suit different environments and acquisition channels.

References

[1] Adami, A., L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan and S. Sivadas, "QUALCOMM-ICSI-OGI Features for ASR", in *Proceedings of International Conference on Spoken Language Processing* , Denver, Colorado, USA, Sep, 2002

[2] Athineos, M. and D. Ellis, "Frequency-domain linear prediction for temporal features", *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, St. Thomas, US Virgin Islands, 2003

[3] Atick, J.J. "Could information theory provide an ecological theory of sensory processing?", in *Network: Computation in Neural Systems*, Vol. 3, 213-251, 1992

[4] Bilmes, J. "Maximal mutual information based reduction strategies for cross-correlation based joint distributional modeling", *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, SP14.6, Seattle, 1998

[5] Brown, P. "The Acoustic-Modeling Problem in Automatic Speech Recognition"*,* PhD Thesis, *Computer Science Department, Carnegie Mellon University*. 1987.

[6] Bourlard, H. and Ch. Wellekens, "Links Between Markov Models and Multilayer Perceptrons.", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 12: 1167-1178, 1990

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

[8] Cooper, F.S., A.M. Liberman, J.M. Borst, "The Interconversion of Audible and Visible Patterns as a Basis for Research in the Perception of Speech", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 37, No. 5, 318-325, 1951

[9] Depireux, D.D., J.Z. Simon, D.J. Klein, S.S. Shamma, "Spectro-Temporal Response Fields Characterization with Dynamic Ripples in Ferret Primary Auditory Cortex", in *J. Neurophysiology,* Vol. 85, 1220-1234, 2001

[10] Davis, S.B. and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. on Acoustics, Speech, and Signal Processing,* Vol. 28, 357-366, 1980

[11] deCharms, C.R., D. Blake, M.M. Merzenich , “Optimizing sound feature for cortical neurons”, *Science,* Vol. 280, May 29, 1998

[12] Fletcher, H. “Speech and hearing in communication”*, The ASA edition,* edited by J.B. Allen, Acoust. Soc. Am., 1995, reissue of the original edition from 1953

[13] Fukunaga, K. “Statistical Pattern Recognition”, *Academic Press*, San Diego, 1990.

[14] Furui, S. “Cepstral analysis technique for automatic speaker verification”, *IEEE Trans. on Acoustic, Speech, & Signal Processing,* vol. 29, 254-272, 1981

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

[16] Hermansky, H. “Perceptual linear predictive (PLP) analysis of speech”, *J. Acoust. Soc. Am.*, vol. 87, no. 4, 1738-1752, 1990.

[17] Hermansky, H. "The Modulation Spectrum in the Automatic Recognition of Speech", *Workshop on Automatic Speech Recognition and Understanding Proceedings,* S. Furui, B.H. Juang, W. Chou, Eds. IEEE 1997

[18] Hermansky, H. and S. Sharma, "TRAPS Classifiers of Temporal Patterns", in *Proceedings of International Conference on Spoken Language Processing 1998*, Sydney, Australia, 1998,

[19] Hermansky, H. and P. Jain,”Band-independent speech event categories for TRAP based ASR”, *Proc. Eurospeech 2003,* Geneva 2003

[20] Hermansky, H. "Should recognizers have ears?", in *Speech Communication*, vol. 25, 3-27, 1998

[21] Hermansky, H. and D.P.W. Ellis and S. Sharma, "Connectionist Feature Extraction for Conventional HMM Systems", in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, 2000

[22] Hermansky, H. and S. Sharma, "Temporal Patterns (TRAPS) in ASR of Noisy Speech", in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Phoenix, Arizona, USA, Mar, 1999

[23] Hermansky, H. and N. Malayath, "Spectral Basis Functions from Discriminant Analysis", in *Proceedings of International Conference on Spoken Language Processing*, Sydney, Australia, 1998

[24] Hermansky, H and N. Morgan, "*RASTA processing of speech*", IEEE Trans. Speech and Audio Processing, vol.2, no.4, pp.578-589, 1994.

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

[26] www.icsi.berkeley.edu/speech/faq/ICSI_SPEECH_FAQ

[27] Jain, P., *Temporal Patterns of Frequency-Localized Features in ASR*, PhD. thesis, Department of Electrical and Computer Engineering, OGI School of Oregon Health & Sciences University, Portland, Oregon, 2003

[28] Jain, P. and H. Hermansky, “Effect of combining temporal patterns from critical-bands on ASR “, *Proc. Eurospeech 2003*, Geneva 2003

[29] Jelinek F., *Statistical Methods for Speech Recognition*, MIT Press, 1998

[30] von Kempelen, W. *Mechanismus der menschlischen Spraeche,* Vienna, J.B. Degen, 1791

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

[32] Kajarekar, S. and H. Hermansky, " 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, 2000

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

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

[35] Malayath, N. and H. Hermansky, "Bark resolution from speech data", *Proceedings International Conference on Spoken Language Processing* 2002, Denver, Colorado, September 2002.

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

[37] Marr, D., *Vision*, W.H. Freeman, San Francisco, 1982

[38] Mermelstein, P., "Distance measures for speech recognition, psychological and instrumental", in *Pattern Recognition and Artificial Intelligence,* R.C.H. Chen, ed., Academic Press: New York, 374-388, 1976

[39] Morgan, N., Qifeng Zhu, A. Stolcke, K. Sonmez, , S. Sivadas,, T. Shinozaki, M. Ostendorf, P. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, O. Cretin, H. Bourlard, M. Athineos, "Pushing the envelope - aside", *IEEE Signal Processing Magazine,* 22 (5), 81- 88, 2005

[40] Sachs, M. and E. Young, "Encoding of steady state vowels in the auditory nerve: representation in terms of discharge rate", *J. Acoust. Soc. Am.* 66, 470-479, 1979

[41] Schwarz, P., P. Matejka and J.Cernocký, "Recognition of Phoneme Strings using TRAP Technique", *Proc. Eurospeech 2003*, Geneva 2003

[42] Sivadas, S. and H. Hermansky, "Hierarchical Tandem Feature Extraction", in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* 2002, Orlando, Florida, USA, May, 2002

[43] Umesh, S., L. Cohen and D. Nelson, Frequency warping and speaker normalization, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Munich, Germany, 1997

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

[45] Young, E. and M. Sachs, "Representation of steady-state vowels in the temporal aspects of the discharge patterns of population of auditory nerve fibers, *J. Acoust. Soc. Am* 66, 1381-1403, 1979.

[46] Hermansky: "Human Speech Perception: Some Lessons from Automatic Speech Recognition, Text", *Speech and Dialogue 2001*, Matousek et al. Eds. Springer 2001.

[47] Chen L., Liao H.,Ko M.,Lin J. and Yu G.: “A new lda-based face recognition system which can solve the small sample size problem”. *Pattern Recognition*, 33(10):1713-1726, Oct 2000

[48] H. Yu and J. Yang: “A direct LDA algorithm for high-dimensional data with application to face recognition”, *Pattern Recognition*, vol. 34, pp. 2067-2070, 2001.

[49] van Vuren S., Hermansky H.: Data-Driven Design of RASTA-Like Filters, *Proc. of Eurospeech 87*, Rhodes, Greece, 1997.

[50] Hermansky H. and Fousek P.: “Multi-resolution RASTA filtering for TANDEM-based ASR”, *Proc.of Interspeech 2005*, Lisboa, 2005.

[51] Kajarekar A., Yegnanarayana B. and Hermansky H., ”A Study of Two Dimensional Linear Discriminants for ASR”, *Proc. of ICASSP’01*, Salt Lake City, Utah, USA, May, 2001

[52] Kozhevnikov and L. Chistovich, “Speech, Articulation and Perception”, translated by *Joint Publications Research Service*, Washington, 1965

[53] Jieping Y., Ravi J. and Qi L.: “Two-Dimensional Linear Discriminant Analysis, Advances”, in *Neural Information Processing Systems* 17, MIT Press, Cambridge, MA, pp. 1569-1576, 2005.

[54] Haeb-Umbach, R., and Ney, H. :”Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition, in *Proc. ICASSP 1992*, San Francisco, CA, March 1992, pp I.13-I.16.

[55] Picheny M., Personal communications 2000.

[56] Chistovich, Vencov, Granstrem, i dr. Fiziologija rechi. Vosprijatie rechi chelovekom (Nauka, 1976)

[57] Junqua J-C., “Robustness and cooperative multimodal man-machine communication applications,” SMMD, 1991, vol. I, pp. 101–112.

[58] Ajmera J., McCowan I., and Bourlard H., “Speech/music segmentation using entropy and dynamism features in a hmm classification framework,” *Speech Communication*, vol. 40, pp. 351–363, May 2003.

[59] Scheirer E. and Slaney M., “Construction and evaluation of a robust multifeature speech/music discriminator,” ICASSP, 1997, vol. 1, pp. 1331–1334.

[60] Drullman R., Festen J., and Plomp R., “Effect of reducing slow temporal modulations on speech reception,” *Journal Acoust. Soc.*, vol. 95, pp. 2670–2680, 1994.

[61] Kanedera N., Arai T., Hermansky H., and Pavel M., “On the relative importance of various components of the modulation spectrum for automatic speech recognition,” *Speech Communications*,

vol. 28, pp. 43–55, 1999.

[62] Hermansky H., “Auditory modeling in automatic recognition of speech,” ECSAP, 1997.

[63] Rabiner L.R., and Sambur M.R., “An algorithm for determining the endpoints of isolated utterances,” *Bell Systems Tech. Jour.*, vol. 54, pp. 297–315, February 1975.

[64] Young S. and et al, *The HTK Book Version 2.2*, Entropic Ltd, 1999.

[65] Dines J., Vepa J., and Hain T., “The segmentation of multichannel meeting recordings for automatic speech recognition,” ICSLP, 2006.

[66] Lincoln M., McCowan I., Vepa J., and Maganti H., “The multichannel wall street journal audio-visual corpus (mc-wsj-av): Specification and initial experiments,” ASRU, 2005.

[67] McCowan I., Maganti H., Gatica-Perez D., and et al, “Speech acquisition in meetings with an audio-visual sensor array,” ICME, 2005.

[68] J. Makhoul, “Spectral linear prediction: Properties and applications,” in *Trans. ASSP*, vol. 23:3, Jun 1975, pp. 283–296.

[69] H. Hermansky, H. Fujisaki, and Y. Sato, “Analysis and synthesis of speech based on spectral transform linear predictive method,” in *Proc. ICASSP*, vol. 8, Apr 1983, pp. 777–780.

[70] R. Drullman, J. M. Festen, and R. Plomp, “Effect of temporal envelope smearing on speech reception,” *The Journal of the Acoustical Society of America*, vol. 95, 1994.

[71] H. Hermansky, “TRAP-TANDEM: Data-driven extraction of temporal features from speech,” in *Proc. IEEE ASRU*, St. Thomas, USVI, 2003.

[72] J. Herre and J. Johnston, “Enhancing the Performance of Perceptual Audio Coders by Using Temporal Noise Shaping (TNS),” in *Proc. 101st AES Conv.*, Nov 1996.

[73] J. L. Marple, “Computing the discrete-time ’analytic’ signal via fft,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 47, Sep 1999.

[74] P. Jain and H. Hermansky, “Beyond a single criticalband in TRAP based ASR,” in *Proc. Eurospeech*, Geneva, Switzerland, Nov 2003.

[75] R. Kumaresan and A. Rao, “Model based approach to envelope and positive instantaneous frequency estimation of signal with speech applications,” *The Journal of the Acoustical Society of America*, vol. 105, 1999.

[76] Makhoul J., “Linear Prediction : A Tutorial Review”, in Proc. of IEEE, Vol. 63, No. 4, April 1975.

[77] OppenheimA.V., SchaferR.W.,“Discrete-Time SignalProcessing”,2ndEd.,Prentice-Hall,NJ, USA, 1998.

[78] Churchill R. V., Brown J. W., “Introduction to Complex Variables Applications”, 5th Ed., McGraw-Hill Book Company, NY, USA, 1982.

[79] Motlicek P., Hermansky H., Garudadri H., and Srinivasamurthy N. “Speech Coding based on Spectral Dynamics”, in "Ninth International Conference on TEXT, SPEECH and DIALOGUE (TSD), 2006.

[80] Motlicek P., Ullal V., Hermansky H., “Wide-Band Perceptual Audio Coding based on Frequency-Domain Linear Prediction”, accepted for IEEE ICASSP conference 2007.

[81] Athineos, M., H. Hermansky, D. P.W. Ellis. “LP-TRAP: Linear predictive temporal patterns”, in Proc. of ICSLP, pp. 1154-1157, Jeju, S. Korea, October 2004.