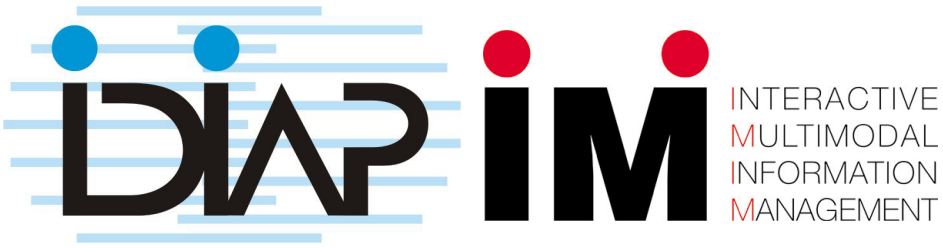

IDIAP RESEARCH REPORT

Some Emerging Concepts in Speech Recognition

Hynek Hermansky[a] and Hervé Bourlard[a,b]

IDIAP-RR -03-82

DECEMBER, 2003

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

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

[a] IDIAP, Martigny, Switzerland
[b] EPFL, Lausanne, Switzerland.

ABSTRACT

The paper presents a work-in-progress on several emerging concepts in Automatic Speech Recognition (ASR), that are being currently studied at IDIAP. This work can be roughly categorized into three categories: 1) data-guided features, 2) features based on modulation spectrum of speech, 3) minimum entropy based multi-stream information fusion.

1. DATA-GUIDED FEATURES

Summary: *Optimal set of features for classification are posteriors of classes. Non-linear classifier, trained on labeled development data, is used to estimate posterior probabilities of sub-word unit classes. These estimates are gaussianized and whitened and subsequently used as features for the conventional HMM-based ASR.*

Typical features for ASR such as Mel Cepstrum or PLP represent short-term spectral envelope, warped in frequency to emulate non-equal spectral sensitivity of hearing, smoothed by some means (cepstral truncation or autoregressive modeling), with logarithmically compressed amplitude axis (to ensure more Gaussian distribution) and approximately decorrelated by a projection on cosine basis. All the above described operations are *ad hoc* but they are justified with respect to the subsequent classification.

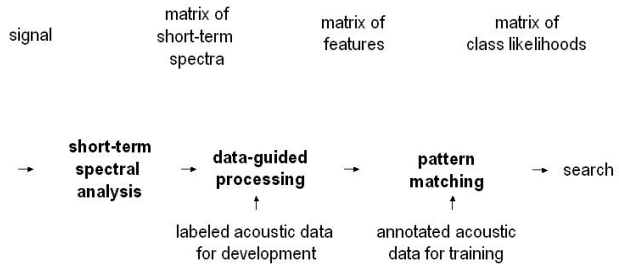

Fig. 1 Data-guided feature extraction. Unlike the conventional feature extraction, the additional processing of the short-term spectrum of speech is derived by optimizing some criteria using labeled development database.

An alternative strategy is to derive at least some stages of the feature extraction by optimizing the feature set performance on the target application or at least on some task that is related to the target application. Most of ASR systems employ phoneme-related elements as sub-word units. Features that would ensure optimal separation among phonemes would therefore make sense. When the speech database labeled by phoneme classes is available, it is possible to derive by a linear discriminant analysis (LDA) a projection that optimizes so called Fisher ratio (ratio between-class variance and within-class variance). Applying LDA to short-term spectral features yields basis that exhibit auditory-like non-equal spectral sensitivity [Malayath and Hermansky]. Applying it to temporal vectors of critical-band spectral energies yields FIR filters that suppress modulation spectrum components below 1 Hz and above 15 Hz [van Vuuren and Hermansky].

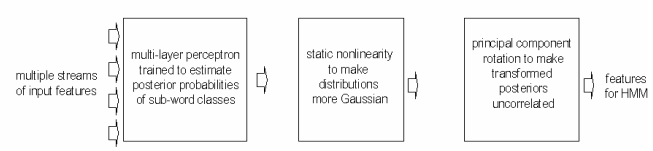

Fig. 2 TANDEM feature extraction. Similarly as in the ANN/HMM hybrid system, any available evidence in the data can be used for deriving posterior probabilities of sub-word classes. In contrast to the conventional ANN/HMM, these posterior probabilities are processed through a static nonlinearity to make them more Gaussian distributed, and whitened by a PCA transform (derived on some development data) and used as a featured in the conventional GMM-based HMM system.

A non-linear extension of the LDA analysis is a multi-layer perceptron (MLP), trained on labeled data to estimate posterior probabilities of sub-word units [Bourlard and Morgan]. These estimates can subsequently be gaussianized and whitened to be used as features for the conventional HMM-based ASR [Hermansky et al.]. One of advantages of the TANDEM technique is that it can use arbitrary distributed and highly-correlated data and to combine various types of data. Further, the discriminative training of the MLP could prove advantageous in the maximum-likelihood based HMM systems.

The performance of the recognizer typically improves with increasing amount of the acoustic data that is used in the training of the recognizer. We have some evidence that the performance of the recognizer with the data-guided feature extraction module increases with the increase of amounts of either the training or the development data [Sivadas and Hermansky, ICASSP-04]. One of the results is shown in the Fig. That shows the error of the recognizer (OGI digits) as a function of the amount of the training data. As seen, the recognizer with the data-driven TANDEM feature extraction (derived on about 3 hours of OGI Stories data) degrades much slower with the decreasing amount of the target training data than the recognizer with the conventional PLP feature extraction. That offers an interesting compromise where the data-guided feature extraction module derived on one batch of the development data could be used with advantage on a different task with smaller amounts of the training data.

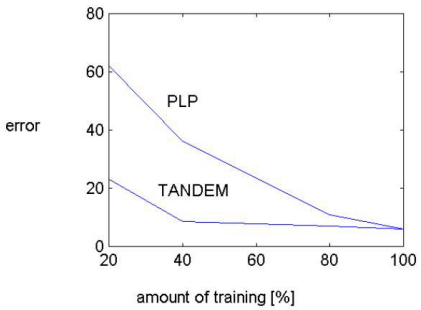

Fig. 3 Performance of a conventional (PLP-based) ASR and the TANDEM-based ASR as a function of the amount of the training data. The conventional system shows the often observed degradation of its performance with decreasing amount of the available training. The TANDEM-based system (with its front-end trained on an additional large database unrelated to the task) exhibits much slower degradation of its performance.

2. FEATURES BASED ON MODULATION SPECTRUM OF SPEECH

Summary: *Information is carried in changes of the signal. Modulation spectrum reflects the rate of change of components of spectral envelope of speech. Two techniques that use the modulation spectrum are discussed*

1. *Projection of a segment of temporal trajectory of cepstral features on sine and cosine bases as additional (dynamic) features in ASR.*
2. *Frequency-localized posterior probabilities are derived from relatively long (300-1000 ms) temporal trajectories of spectral densities and used as an input to the TANDEM feature extraction.*

The modulation spectrum can be derived by spectral analysis of temporal trajectories of short-term spectral envelopes. The modulation spectrum can be modified by applying filters to temporal trajectories of speech features [Hermansky and Morgan]. Since the modulation frequencies of interest are as low as 1 Hz, relatively long segments of temporal trajectories need to be used to derive the modulation spectrum of speech.

2.1. Spectral analysis of modulation spectrum

Spectral decomposition of modulation spectrum has been suggested by Kay [Kay and Matthews] and later studied by Hougast [Houtgast] and Dau and his colleagues [Dau et al]. Kanedera and his colleagues [Kanedera et al] has shown relative importance of various components of the modulation spectrum for speech intelligibility and for ASR. Further, they studied multi-stream ASR which employed several sub-streams with different modulation spectrum components, derived by projecting 400 ms segments of temporal trajectories of PLP cepstral coefficients on three sine and cosine basis (thus effectively emulating modulation filter-bank with 2.5, 5.0 and 7.5 Hz center frequencies . Use of cosine transform was also investigated. Consistently with use of conventional dynamic features for ASR, outputs from all filters were concatenated into one long feature vector that formed the input to the subsequent pattern classification.

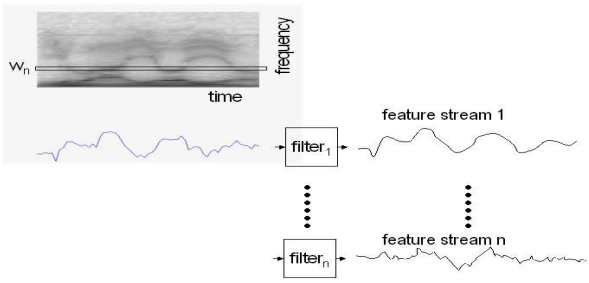

Fig. 4 Decomposition of the modulation spectrum of speech. The modulation spectrum is the spectrum of temporal trajectories of the critical-band logarithmic speech spectrum [Houtgast and Steeneken]. This concept can be generalized to other feature sets (such as e.g. cepstral coefficients). By passing a particular time trajectory through a bank of filters, the filter outputs provide new sets of features that represent different parts of` the modulation spectrum of the original feature space.

More recently, Tyagi and his colleagues [Tyagi et al] used this technique on somehow shorter segments (110 ms) of temporal trajectories of Mel cepstral coefficients and called the technique Mel Cepstrum Modulation Spectrum (MCMS).

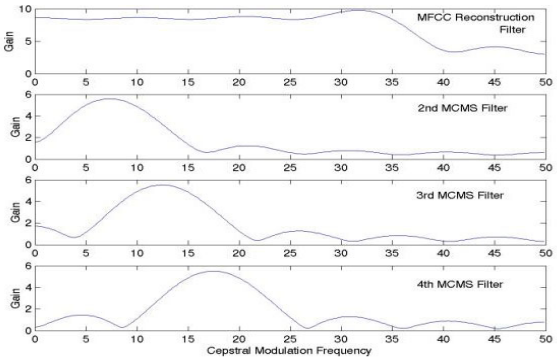

Fig. 5 The frequency characteristics of modulation filters applied in [Tyagi et al]. The first filter represents low-pass that suppresses modulation spectrum components higher than about 35 Hz. The higher ones enhance the rate of change around 8 Hz, 13 Hz and 17 Hz respectively.

Frequency characteristics of modulation spectrum filters that were applied are shown in Fig. As seen, the filters are bandpass with progressively higher center frequencies at around 6, 12, and 18 Hz. The first filter is a low-pass with a cutoff frequency around 30 Hz.

2.2. TRAP-TANDEM

The above described techniques typically operate on cepstral trajectories (rather than on the trajectories of speech spectrum) and therefore use the whole spectral envelope. Another way is to directly use temporal trajectories of spectral energies from individual critical band filters. This in effect results in using frequency-localized modulation spectra at individual carrier frequencies. Such an approach then has advantages of multi-band techniques in allowing for alleviating features from unreliable frequency bands.

The input to the TRAP estimator is formed from 1-3 time trajectories of critical-band energies. These temporal patterns may be parameterized. The most successful parameterization has been so far the cosine transform (closely resembling PCA (Karhunen-Loeve) transform (KLT) derived on development data). The cosine transform typically allows for at least 50% reduction of dimensionality of the input data. Some benefits are seen when more than one time trajectory is used as an input. In that case, the individual trajectories are concatenated to form a longer input vector, and some form of dimensionality reduction is typically applied. The PCA analysis of the data suggests in this case, that for preserving most of the variability in the multiple-trajectory data, the data from the individual trajectories should be averaged and differentiated, in effect crudely describing the spectral shape of the local time-frequency pattern in the vicinity of the frequency of interest [Jain and Hermansky 2003, Grezl and Hermansky 2003].

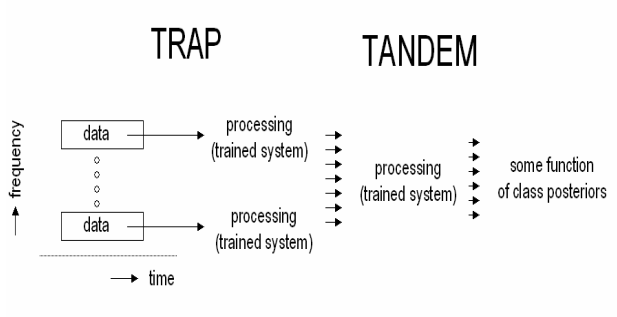

Fig. 6 Schematic picture of the complete TRAP-TANDEM system. The TRAP part takes rather long (up to 1 s) and relatively narrow (1-3 Bark) patches from the critical-band speech spectrum at different frequencies and (using trained MLPs) computes scaled likelihoods of speech events of interest . The TANDEM part takes the concatenated vectors of the likelihoods at different frequencies and (using another trained MLP) first converts them to a vector of likelihoods of sub-word units (phonemes). These likelihoods, transformed through a static nonlinearity to make their distribution more gaussian distributed and whitened through the PCA rotation to make them decorrelated, are used as features for the conventional HMM recognizer.

The estimator itself is an MLP that is trained on labeled development data to estimate posterior probabilities of some classes of interest. These classes were initially context-independent phonemes, later collapsed into 6 broad phonetic classes. Different estimators were trained for each of critical band. The current efforts are directed towards deriving one single estimator to be applied at all frequency bands. The classes of interest are then obtained by clustering all mean temporal patterns of critical-band spectral energies (regardless of the frequency location of the particular critical band) within boundaries of the individual phonemes [Hermansky and Jain, Eurospeech 2003].

Outputs from all individual estimators are concatenated into one vector that typically form and input to the TANDEM data-guided feature module (described above). The TRAP-TANDEM scheme is already quite competitive on its own with the more conventional frame-based features. It typically yields noticeable advantage in a multi-stream combination with the conventional front end [Hermansky 2003].

3. MINIMUM ENTROPY BASED INFORMATION FUSION

Summary*: We are investigating here "optimal" ways to recombine multi-stream ASR in which a number of likelihood/posterior estimation processes is being carried out in parallel. In the effective information fusion, the reliable sub-streams would be used and the unreliable ones would be selectively and adaptively suppressed. In the present approach, we investigate the possibility of combing multiple sources of information based on the individual classifier/channel's entropy (associated with individual sub-streams). Low-entropy classifiers/channels are weighted more heavily than high-entropy ones, and the final likelihood/posterior distribution is obtained by fusion of likelihoods/posteriors from the individual sub-stream processes.*

We believe that entropy can be an effective measure to achieve the above goal. Indeed, reliable classification is typically characterized by some classes having much higher likelihood than the other

classes. The unreliable classification, on the other hand, typically yields similar and mediocre likelihoods for all classes. This observation can be quantified by computing the entropy of the classifier's output distribution, which would be low for the reliable estimates and high for the unreliable ones. In the fusion module, the outputs from the individual sub-stream estimators can be then weighted by a function of the inverse of the entropy of the outputs from the sub-stream estimators.

Thus, assuming that the entropy of the MLP output distribution (then associated with posterior probability distribution) is representative of the confidence in the classification, and based on the fact that the ultimate goal of a speech recognition system is to reduce the conditional entropy (conditional on lexical and grammatical constraints), the output vectors of the different MLP classifiers exhibiting low entropy of their output distributions should be weighted more heavily than the high-entropy ones.

In [Misra et al, ICASSP 2003], different recombination schemes based on inverse-entropy weighting were studied.
One weighting scheme that was found useful uses average entropy from all sub-streams as an adaptive threshold. Outputs of all sub-streams that were above this threshold are severely de-weighted (practically eliminated) in the combination, the ones below the threshold were weighted by inverse of their respective entropies. This was shown resulting in significant error reduction, as well as a consistent reduction of the entropy of the resulting (recombined) output distribution.

More recently, a similar combination scheme was also experimentally evaluated in fusing the information from two sub-streams, the one (based on a conventional power spectrum) yielding high performance on clean speech and the other (with a spectrum with enhanced peaks) somehow inferior in clean conditions but more robust to noise [Ikbal et al, submitted to ICASSP 2004]. The inverse entropy based combination of these two sub-streams yielded the feature extraction that was both well performing on the clean speech and was robust to noise.

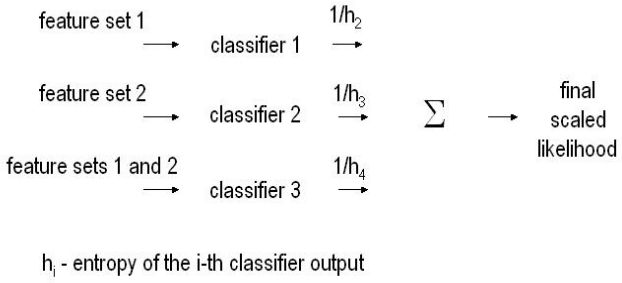

Fig. 7 Information fusion of two information sub-streams using the inverse entropy rule. Three probability estimators are trained on features from two sub-streams, the first two estimators using only the features from the respective individual sub-streams, the third one using the concatenated feature vectors from both sub-streams. The fourth estimator having a null input (thus only using the prior likelihoods) can be also used but is usually omitted. During the recognition, entropies of all estimator outputs are computed for each input frame and the final likelihood estimate is created as a weighted sum of the likelihoods from the individual estimator, where the weighting factors are given by inverse of the entropies of the outputs from the individual estimators. Extensions to more than 2 sub-streams are trivial and have been also investigated.

ACKNOWLEDGEMENTS

The paper summarizes work of many of our colleagues, most of it acknowledged through the references to their original works. Most of the research and the preparing of this paper were supported by Swiss NCCR on Interactive Multimodal Information Management (IM2), by EC grant M4, and by DARPA EARS program.

References

N. Malayath and H. Hermansky: Data-driven spectral basis functions for automatic speech recognition, Speech Communication 40, pp. 449-466, 2003

S. van Vuuren and H. Hermansky: Data-driven design of RASTA-like filters, Proc. EUROSPEECH 1997, pp. 409-412, 1997

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

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

S. Sivadas and H. Hermansky, On use of task-independent training data in TANDEM feature extraction, submitted to ICASSP 2004

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

R.H. Kay and D.R. Matthews, On existence in human auditory patways of channel selectivity to modulation present in frequency modulated tones, J. Physiol. 225, pp. 657-667, 1972

T. Houtgast Frequency selectivity in amplitude-modulation detection, J. Acosut.Soc. Am. 45, pp. 1676-1680, 1989

T. Dau, J. Verhey and A. Kohlrausch, Intristic envelope fluctuations and modulation-detection thresholds for narrowband noise carriers, J. Acoust. Soc. Am. 102, pp. 2906-2919, 1999

N. Kanedera, H. Hermansky and T. Arai, Desired characteristic of modulation spectrum for robust, 1998 automatic speech recognition, Proc. ICASSP 98, Seattle

T. Houtgast and H.J.M. Steeneken, The modulation transfer function in rooms acoustics as a predictor of speech intelligibility, Acustica 28, pp. 66-73, 1973.

V. Tyagi,, I . MacCowan, H. Misra qand H. Bourlard, Met cesptrum modulation spectrum (MCMS) features for robust ASR, Proc. ASRU-2003, St. Thomas, US Virgin Islands, December 2003

P. Jain and H. Hermansky, Beyond a single critical-band in TRAP based ASR, Proc. Eurospeech 2003, Geneva, Switzerland, September 2003.

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

H. Hermansky, TRAP-TANDEM: Data-driven extraction of temporal features from speech, Proc ASRU-2003

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

H. Misra, H. Bourlard and V. Tyagi, New entropy based combination rules in HMM/ANN multi-stream ASR, Proc. ICASSP-2003, pp 741-744, 2003

S. Ikbal, H. Misra, H. Bourlard and H. Hermansky, Phase autocorrelation (PAC) features in entropy based multi-stream for robust speech recognition, submitted to ICASSP-2004.