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: D 2.4
Results of Processing with Advanced Hierarchical Model Including Results on an Accepted ASR Task

Date of deliverable: 30.06.2007
Actual submission date: 06.08.2007

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)	

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.)

D2.4 Results of Processing with Advanced Hierarchical Model Including Results on Accepted ASR Task

IDIAP Research Institute (IDIAP)

Abstract:
The modulation spectrum is a useful representation of a signal for incorporating dynamic information. In this work we investigate how to process different modulation spectrum frequencies from an ASR perspective. Parallel and hierarchical approaches are investigated. Parallel processing combines output of independent classifiers trained on different modulation frequency channels. Hierarchical processing uses different modulation frequencies at different level of the model in a sequential fashion. Experiments are run on a meeting transcription LVCSR task and results are reported on the RT05 evaluation data. Processing modulation frequencies channels with different classifiers provide a consistent reduction in WER. Furthermore hierarchical processing outperforms parallel processing. This model is consistent with several perceptual and physiological studies on auditory processing.

Table of Content

1. Introduction

Conventional speech recognition features are based on short time Fourier transform(STFT) of the signal thus information about speech dynamics are lost. For this reason dynamic features are introduced (e.g. delta features) although they offer a very rough approximation of the temporal changes in the signal. An alternative consist in the study of long segments of critical band energies obtained by STFT i.e. the modulation spectrum of the signal. This approach has been shown to provide useful information on the speech dynamics and to increase noise robustness (e.g. [1],[2]).

Previous studies showed the contributions of different parts of the modulation spectrum for word recognition and suggested the use of multiple resolution analysis ([3]). Gabor filters [4] and multi-resolution RASTA filters [5] have been proposed for extracting different modulation frequencies. Those techniques are generally used into conventional HMM/GMM systems using a TANDEM [6] approach. The multi-resolution representation of the speech signal is used as input into a Neural Network in order to estimate phoneme posterior probabilities then a Log/KLT transform is applied to posterior probabilities that are used as features into conventional ASR systems.

However all previous studies have considered the problem using a single classifier i.e. all modulation frequency channels are processed simultaneously by the same classifier.

Several perceptual [7] and physiological studies [8] suggest that in the auditory system, this processing is done separately for each channel. In this work we investigate from an ASR perspective if different modulation frequencies should be processed in parallel fashion or in sequential (hierarchical) fashion.

Parallel processing uses an independent classifier for each modulation frequency channel. Phonemes posterior distributions are then recombined using a merger classifier. This is somehow similar to multi-band processing [9]. On the other hand, hierarchical processing uses a hierarchy of classifiers that incorporates different modulation frequency at different level in a sequential fashion. Hierarchical classifiers are very popular in the field of computer vision and recently some studies have been proposed on their application to simple phoneme recognition task [10].

In contrary to previous related works on multiple resolution modeling, we study here the use of those two different approaches in a LVCSR task for transcription of meetings.

Training data consists in 100 hours of meetings and results are reported on RT05 evaluation data. The paper is organized as follows: in section 2 we describe multiple resolutions RASTA filtering (MRASTA), in section 3 we describe data and system used for experiments, in sections 4 and 5 we describe respectively parallel and hierarchical processing of modulation frequency channels with results on RT05 evaluation data and in section 6 we discuss conclusions on this work.

2. MRASTA Processing

MRASTA filtering [5] has been proposed as extension of RASTA filtering trough the use of a two dimensional band-pass filter. Critical band auditory spectrum is extracted from short time Fourier transform of a signal every 10 ms. A one second long temporal trajectory is filtered with a bank of low-pass filters, represented by eight first derivatives (G1) and eight second derivatives (G2) of Gaussian functions with variance varying in the range 8-130 ms (see figure 1 - for details see [5]).

Filters are used for all bands. In contrary to [5] we use in this work only 6 filters. In the modulation frequency domain, they correspond to a filter-bank with equally spaced filters on a logarithmic scale. For instance figure 3 plots a critical band auditory spectrogram filtered

with the first, the fourth and the eight Gaussian derivative filters (G1). Those filters provide a multiple-resolution view of the time-frequency plane.
Subsequently frequency derivatives are introduced with a context of three-Bark frequency. This results in a bank of spectro-temporal filters as in figure 2. MRASTA processing is consistent with a large number of studies on auditory processing i.e. perceptual studies on modulation frequencies [7] and spectro-temporal receptive fields (STRF) [11].
MRASTA filtering is used as pre-processing step of the auditory spectrogram for TANDEM system. A Neural Network is used for deriving posterior probabilities of phonetic targets. Phoneme posterior probabilities are then transformed according to TANDEM scheme [6] and used as features in conventional HMM based systems. In next section we describe the ASR system used in this work.

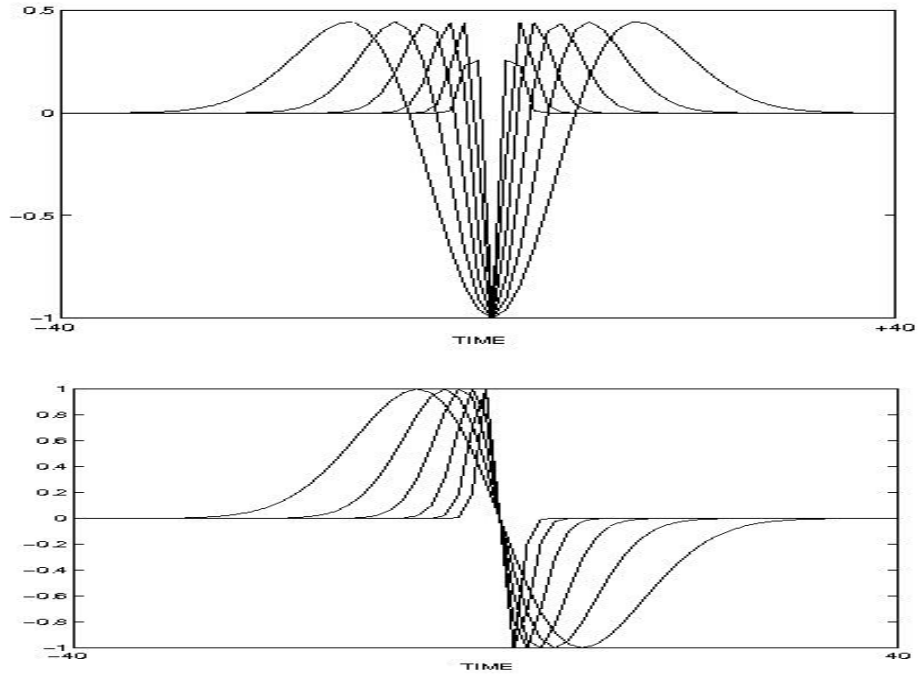

Figure 1. Set of temporal filter obtained by first (G1) and second (G2) order derivation of Gaussian function.

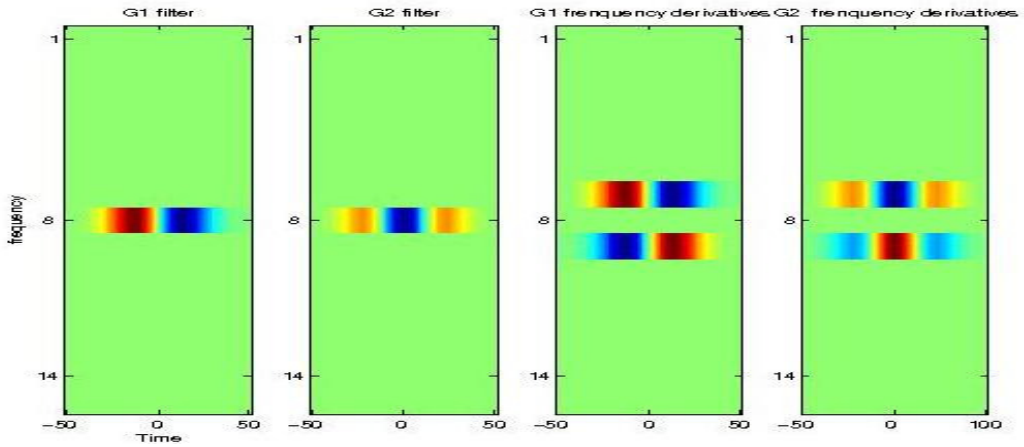

Figure2. Spectro-temporal filters used in MRASTA processing.

3. System description

Experiments are run with the AMI LVCSR system for meeting transcription described in [12]. The training data for this system comprises of individual headset microphone (IHM) data of four meeting corpora; the NIST (13 hours), ISL (10 hours), ICSI (73 hours) and a preliminary part of the AMI corpus (16 hours). Acoustic models are phonetically state tied triphone models trained using standard HTK maximum likelihood training procedures. The recognition experiments are conducted on the NIST RT05s [13] evaluation data. We use the reference speech segments provided by NIST for decoding. The pronunciation dictionary is same as the one used in AMI NIST RT05s system [12]. Juicer large vocabulary decoder [14] is used for recognition with a pruned trigram language model.
TANDEM-MRASTA features are obtained from the all training set. Table 1 reports results for the PLP system and the MRASTA-TANDEM system.

Features	***TOT***	***AMI***	***CMU***	***ICSI***	***NIST***	***VT***
PLP	42.4	42.8	40.5	31.9	51.1	46.8
MRASTA	45.8	47.6	41.9	37.1	53.7	49.7

Table 1. RT05 WER for Meetings data. Baseline PLP system and MRASTA features.

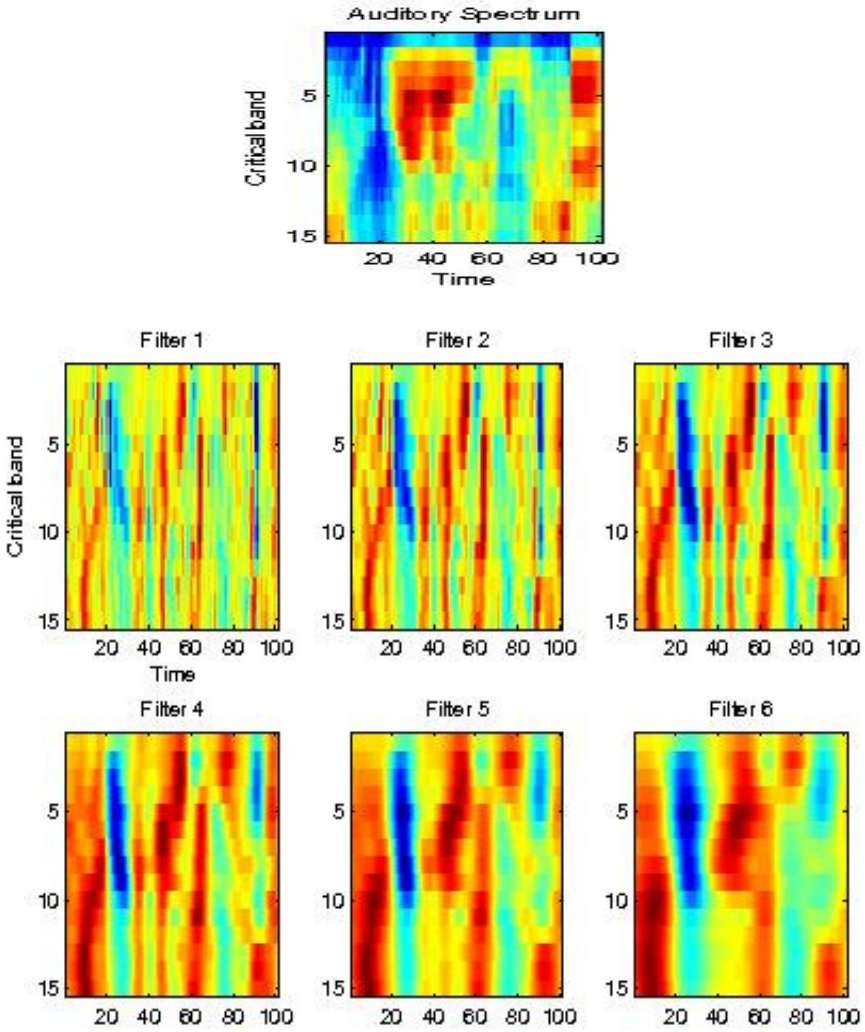

Figure 3. Critical band auditory spectrogram filtered with six Gaussian derivatives filters.

4. Parallel Processing

Let us assume that modeling in each modulation frequency channel is independent; a separate classifier can be trained for each channel and outputs from classifiers recombined into a single posterior stream. This processing would be equivalent to multi-band approach into frequency domain [9]. The filter banks G1 and G2 (6 filters each) are split into two filter banks of 3 filters : the first set (referred as F-high) composed of short filters that capture high modulation frequencies while the second one (referred as F-low) composed of long filters that capture low modulation frequencies. Two independent neural networks are trained on output of filterbanks F-high and F-low and the output is recombined using a neural network merger classifier. The merger neural network takes as input 9 consecutive frames from previous neural networks.
The procedure is depicted in figure 6. Final posterior distributions are used into LVCSR system according to TANDEM approach. Table 2 reports results for high and low modulation frequencies and for combination of high/low frequencies.

Features	***TOT***	***AMI***	***CMU***	***ICSI***	***NIST***	***VY***
High	45.9	48.7	41.9	37.3	53.3	49.2
Low	50.0	51.9	47.6	40.7	57.5	53.1
Combination	41.4	42.7	38.3	32.5	47.4	47.1

Table 2. RT05 WER for high, low modulation frequencies and combination

Out of the two filter-banks, F-high (high modulation frequencies) largely outperforms F-low (low modulation frequencies). Combination of high and low modulation frequencies using a merger classifiers reduces by 4.4% WER w.r.t. the single classifier approach and by 1% absolute w.r.t. the PLP baseline.
This experiment shows that independent processing of different modulation frequencies channels can significantly reduce overall WER w.r.t. single Neural Network classifier.

5. Hierarchical Processing of Modulation Frequencies

In [15], we showed how to incorporate different features through a hierarchical structure: this work can be extended to incorporate low and high modulation frequencies. The proposed system is described in figure 7. Modulation spectrogram is first filtered trough a first filter bank presented as input to a Neural Network that is trained to obtain phoneme posteriors.
Output from the first NN is then presented into a second NN concatenated with features obtained by processing the modulation spectrogram with a second filter-bank. In such a way, estimates from the first net are modified by second net, using different modulation frequencies. This allow for introducing different features in the system in a sequential fashion. In contrary to parallel processing approaches, the order in which modulation frequencies are presented matters. In table 4 we report WER for features obtained both moving from high to low and from low to high frequencies.

Features	***TOT***	***AMI***	***CMU***	***ICSI***	***NIST***	***VT***
Low to High	45.8	48.3	43.5	37.0	52.5	48.5
High to Low	40.0	40.5	37.3	32.2	47.8	42.9

Table 4. RT05 WER for hierarchical modulation frequencies processing: from low to high and from high to low frequencies

When moving in the hierarchy from low modulation to high modulations performance does not improve. On the other hand, moving from high to low modulations produce a significant reduction into final WER 5.8% w.r.t. single classifier approach. Thus recognition improvement is verified also in case of hierarchical processing but only when the processing moves from high to low modulation frequencies.

This is consistent with physiological experiments in [8] in which it is shown that different levels of auditory processing may attend different rates of the modulation spectrum, the higher levels emphasizing lower modulation frequency rates. To verify that improvements in the previous structure is coming from the sequential processing of modulation frequencies and not simply from a hierarchy of Neural Networks we propose another experiments. Posterior features from the MRASTA net that processes all frequency modulation simultaneously are presented as input to a second NN. The second NN does not use any further information from other frequency representations and simply re-process the posterior features. We also investigate the impact of time-context in the second net.

Features	***TOT***	***AMI***	***CMU***	***ICSI***	***NIST***	***VT***
Hier MRASTA	44.2	46.2	41.9	34.6	51.3	48.1

Table 5. RT05 WER for hierarchical modeling.

Table 5 reports WER on RT05 for those hierarchical features. Hierarchical processing improves performances w.r.t. single net MRASTA of 1.6% absolute. However performances do not reach those of architecture in figure 7. This means that the improvements are actually coming from the sequential processing of modulation frequencies and not from the hierarchical classifier itself.

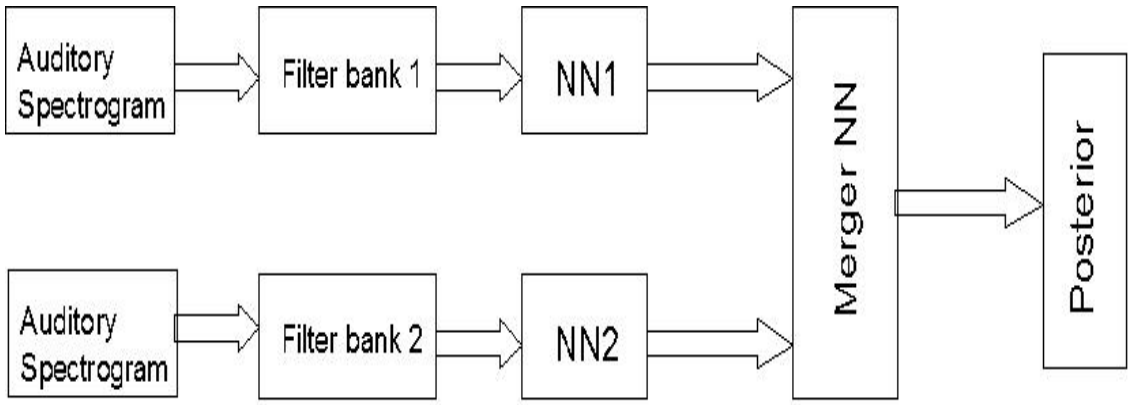

Figure 6. Parallel processing of modulation spectrum frequencies.

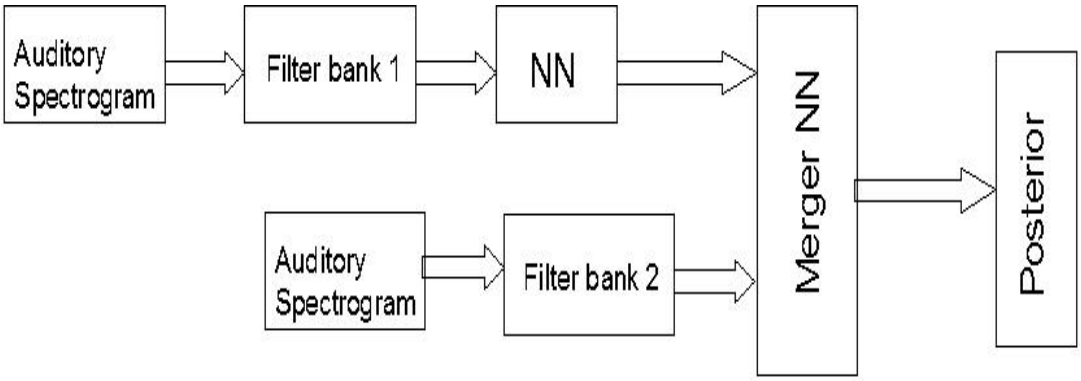

Figure 7. Hierarchical Processing of modulation spectrum frequencies.

6. Summary and Discussions

In this work we discuss parallel and hierarchical processing of different parts of the modulation spectrum domain. Use of modulation frequency filter bank is motivated by perceptual and physiological evidence e.g. [7],[11] and has been proposed in [5] for ASR application. In previous related works, experiments have been conducted from a single classifier perspective.
However studies like [7] and [8] suggests that modulation frequencies channels are processed separately.
We investigate here the processing of different modulation frequencies in both parallel and hierarchical fashion. Experiments are run in a LVCSR task using TANDEM approach. Table 3 summarize results of all previous experiments. Baseline PLP system outperforms single net MRASTA features.

In order to consider high and low modulation frequencies each the G1 and G2 filter-banks are split in two filter-banks of three filters. In parallel processing (see figure 6) two independent Neural Networks are trained on it and outputs combined.
This approach reduces WER of 4.4% absolute w.r.t. the single Neural Network approach.
On the other hand, we investigated the use of hierarchical processing as in figure 2 in which different modulation frequencies are processed in a hierarchical fashion. When the processing order goes from high to low frequencies a 5.8% improvement is obtained while when processing order goes from low to high frequencies, overall WER is similar to the single NN MRASTA. Furthermore High to low frequencies processing outperforms baseline PLP system of 2.4% . In order to verify that the improvement is actually coming from processing different modulation frequencies at different level of the hierarchy we reprocessed MRASTA posteriors with another NN without adding any another information from the time-frequency plane. This further reduces WER of 1.6% but does not achieve recognition rates of architecture in figure 7.

To summarize, separate processing of modulation frequencies improves considerably performances compared to approaches that uses single classifier. Out of the two proposed methods, hierarchical processing is outperforming parallel processing. In those experiments we found that in the hierarchical architecture best performance is obtained when high modulation frequencies are used first and subsequently low frequencies are processed. This is consistent with physiological observation on auditory system [8] in which early auditory stage

emphasizes high modulation frequencies and higher stage emphasizes low modulation frequencies.

Features	*PLP*	*MRASTA*	*Hier MRASTA*	*High*	*Low*	*Comb*	*High to Low*	*Low to High*
WER	42.4	45.8	44.2	50.0	45.9	41.4	40.0	45.8

References

[1] HermanskyH., "Should recognizers have ears?" Speech Communications, vol. 25, pp. 3–27, 1998.

[2] Morgan N. and Greenberg S. Kingsbury B.E.D., "Robust speech recognition using the modulation spectrogram," Speech Communication, vol. 25, pp. 117–132, 1998.

[3] Hermansky H. Kanedera H., Arai T. and Pavel M., "On the importance of various modulation frequencies for speech recognition," Proc. of Eurospeech Eurospeech '97, 1997.

[4] Kleinschmidt M., "Methods for capturing spectrotemporal modulations in automatic speech recognition," Acustica united with Acta Acustica, vol. 88(3), pp. 416–422, 2002.

[5] Hermansky H. and Fousek P., "Multi-resolution rasta filtering for tandem-based asr.," in Proceedings of Interspeech 2005, 2005.

[6] Hermansky H., Ellis D., and Sharma S., "Connectionist feature extraction for conventional hmm systems.," Proceedings of ICASSP, 2000.

[7] Kohlrausch A. Dau T., Kollmeier B., "Modeling auditory processing of amplitude modulation .i detection and masking with narrow-band carrieres.," J. Acoustic Society of America , , no. 102, pp. 2892–2905, 1997.

[8] Miller et al., "Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex," The jounral of Nuerophysiology , vol. 87(1), 2002.

[9] Hermansky H., Tibrewala S., and Pavel M., "Towards asr on partially corrupted speech," Proc. ICSLP 1996.

[10] Rifkin et al., "Phonetic classification using hierarchical, feed-forward spectro-temporal patch based arthitectures," Tech. Rep. TR-2007-007,MIT-CSAIL, 2007.

[11] Kelin D.J. Depireux D.A., Simon J.Z. and Shamma S.A., "Spectro-temproal response field characterization with dynamic ripples in ferret primary auditory cortex," J. Neurophysiol., vol. 85(3), pp.1220–1234, 2001.

[12] Hain T. et al, "The 2005 AMI system for the transcription of speech in meetings," NIST RT05 Workshop, Edinburgh, UK. , 2005.

[13] http://www.nist.gov/speech/tests/rt/rt2005/spring/ ," .

[14] Moore D. et al., "Juicer: A weighted finite state transducer speech coder," Proc.MLMI 2006Washington DC.

[15] Valente F. et al., "Hierarchical neural networks feature extraction for lvcsr system," Proc. of Interspeech 2007, 2007.