IDIAP Research Report

Hierarchical and Parallel Processing of Modulation Spectrum for ASR applications

Fabio Valente [a] and Hynek Hermansky [a]

IDIAP–RR 07-45

January 2008

published in
ICASSP 2008

[a] IDIAP Research Institute, Martigny, Switzerland

IDIAP Research Institute **www.idiap.ch**
Av. des Prés–Beudin 20 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 07-45

Hierarchical and Parallel Processing of Modulation Spectrum for ASR applications

Fabio Valente and Hynek Hermansky

January 2008

published in
ICASSP 2008

Abstract. The modulation spectrum is an efficient representation for describing dynamic information in signals. In this work we investigate how to exploit different elements of the modulation spectrum for extraction of information in automatic recognition of speech (ASR). Parallel and hierarchical (sequential) approaches are investigated. Parallel processing combines outputs of independent classifiers applied to different modulation frequency channels. Hierarchical processing uses different modulation frequency channels sequentially. Experiments are run on a LVCSR task for meetings transcription and results are reported on the RT05 evaluation data. Processing modulation frequencies channels with different classifiers provides a consistent reduction in WER (2% absolute w.r.t. PLP baseline). Hierarchical processing outperforms parallel processing. The largest WER reduction is obtained trough sequential processing moving from high to low modulation frequencies. This model is consistent with several perceptual and physiological studies on auditory processing.

1 Introduction

Conventional speech recognition features are based on short-time Fourier transform (STFT) of short (20-30 ms) segments of speech signal. STFT is able to extract instantaneous levels of individual frequency components of the signal. The information about the spectral dynamics is typically carried in so called dynamic features, representing temporal differentials of the spectral trajectory at the given instant.

An alternative is to use long segments of spectral energy trajectories obtained by STFT i.e. the modulation spectrum of the signal (see [1],[2]). Several studies have been carried out to evaluate the importance of the different parts of the modulation spectrum for ASR applications [3] showing that frequency range in between 1-16Hz with emphasis on 4 Hz is critical for speech recognition. However in those work, modulation frequencies have been studied with uniform resolution.

The use of multiple resolution filter-bank in ASR has been addressed in [4]. Filter-bank consists of a set of multi-resolution RASTA filters (MRASTA) with constant bandwidth on a logarithmic scale and is qualitatively consistent with model proposed in [5]. Other studies that consider multiple resolution modeling with Gabor filters includes [6] and [7]. All those works used a single classifier for the whole range of modulation frequencies.

Some studies suggest processing of modulation spectrum in separate frequency channels. Thus, [8] observes that different levels in the hierarchy of auditory processing emphasize different segments of modulation frequency range, the higher processing level emphasizing lower modulation frequencies.

This paper investigates if there is any advantage in ASR in processing different parts of the modulation frequencies in separate frequency channels. Further we also study if the different parts of the modulation spectrum should be processed in parallel or sequentially (hierarchically). An Artificial Neural Network classifier (NN)(the feed-forward Multi-Layer Perceptron) is applied for estimating phonemes posterior probabilities.

We limit our investigation to only two separate modulation frequency channels that consider respectively high and low frequencies.

The parallel processing uses a separate NN classifier for high and low frequencies. Classifiers outputs are then combined together using a merger neural network in order to provide a single phoneme posterior estimates . This topology is depicted in figure 3.

The hierarchical processing uses a hierarchy of classifiers that incorporates sequentially different modulation frequency bands at different processing levels. This architecture is similar to the one we proposed in [9] for incorporating different feature sets trough a hierarchy of neural networks and it is depicted in figure 4. Hierarchical classifiers are very common in the field of computer vision and recently some studies have been proposed on their application to simple phoneme recognition task [7].

We study the ASR performance on Large Vocabulary Conversational Speech (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 resolution 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 frequencies with results on RT05 evaluation data and in section 6 we discuss conclusions on this work.

2 MRASTA processing

In this section, we describe MRASTA filtering [4] which has been proposed as extension of RASTA filtering. MRASTA filters extract different modulation frequencies using a set of multiple resolution filters.

Feature extraction is composed of the following parts: critical band auditory spectrum is extracted from short time Fourier transform of a signal every 10 ms. A one second long temporal trajectory in each critical band is filtered with a bank of band-pass filters. Those filters represent first derivatives $G1 = [g1_{\sigma_i}]$ (equation 1) and second derivatives $G2 = [g2_{\sigma_i}]$(equation 2) of Gaussian functions with

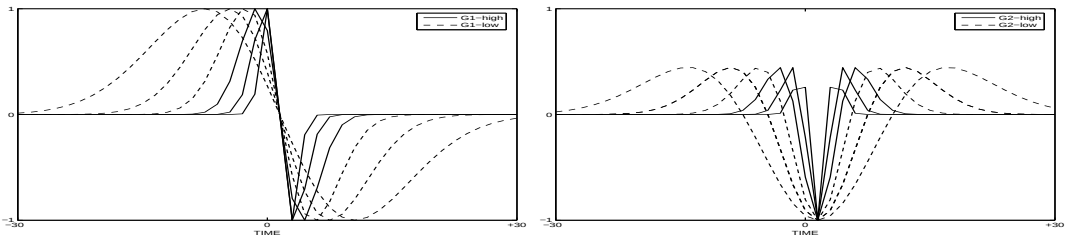

Figure 1: Set of temporal filter obtained by first (G1 left picture) and second (G2 right picture) order derivation of Gaussian function. G1 and G2 are successively split in two filter bank (G1-low and G2-low, dashed line) and (G2-high and G2-high continuous line) that filter respectively high and low modulation frequencies.

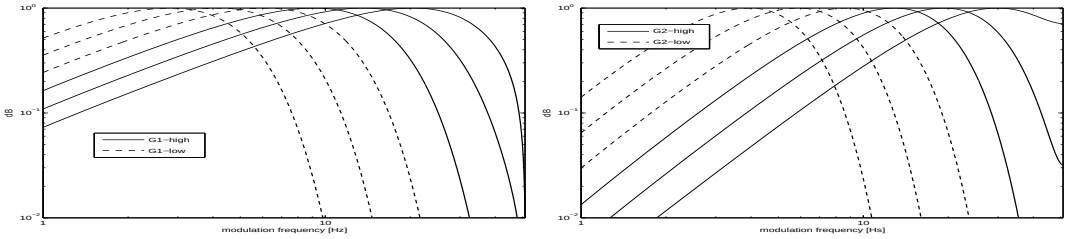

Figure 2: Normalized frequency response of G1 (left picture) and G2 (right picture). G1 and G2 are successively split in two filter bank. G1-low and G2-low (dashed lines) emphasize low modulation frequencies while G1-high and G2-high emphasize high modulation frequencies

variance σ_i varying in the range 8-130 ms (see figure 1). In effect, the MRASTA filters are multi-resolution band-pass filters on modulation frequency, dividing the available modulation frequency range into its individual sub-bands.

$$g1_{\sigma_i}(x) \quad \propto \quad -\frac{x}{\sigma_i^2}\, exp(-x^2/(2\,\sigma_i^2)) \tag{1}$$

$$g2_{\sigma_i}(x) \quad \propto \quad (\frac{x^2}{\sigma_i^4} - \frac{1}{\sigma_i^2})\, exp(-x^2/(2\,\sigma_i^2)) \tag{2}$$

$$\text{with} \quad \sigma_i = \{0.8\,,\, 1.2\,,\, 1.8\,,\, 2.7\,,\, 4\,,\, 6\}.$$

Unlike in [4], filter-banks G1 and G2 are composed of six filters rather than eighth, leaving out the two filters with longest impulse responses. In the modulation frequency domain, they correspond to a filter-bank with equally spaced filters on a logarithmic scale (see figure 2). Identical filters are used for all critical bands. Thus, they provide a multiple-resolution representation of the time-frequency plane. Additionally, local frequency slopes are computed at each critical band by frequency differentiation over the three neighboring critical bands (for details see [4]). Thus the feature vector is composed of 336 components. The resulting multiple resolution representation of the critical-band time-frequency plane is used as input for a Neural Network that estimates posterior probabilities of phonetic targets. Phoneme posterior probabilities are then transformed using TANDEM scheme [10] (i.e. according to a Log/KLT transform) and used as features in conventional HMM based system, described in the next section.

Filter-Banks G1 and G2 cover the whole range of modulation frequencies. We are interested in processing separately different parts of the modulation spectrum and we limit the investigation to two parts. Filter-Banks G1 and G2 (6 filters each) are split in two separate filter bank G1-low, G2-low and G1-high and G2-high that filter respectively high and low modulation frequencies. We define G-high

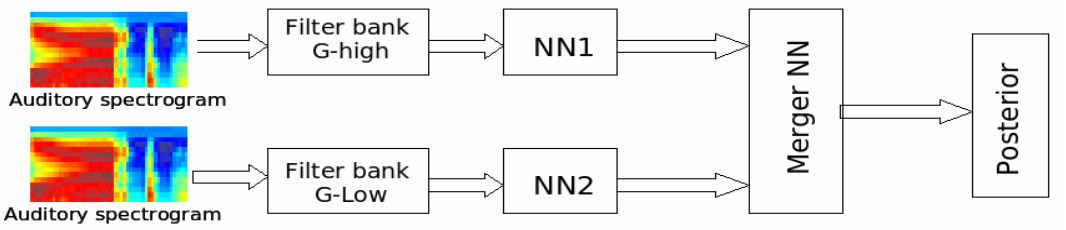

Figure 3: Parallel processing of modulation spectrum frequencies.

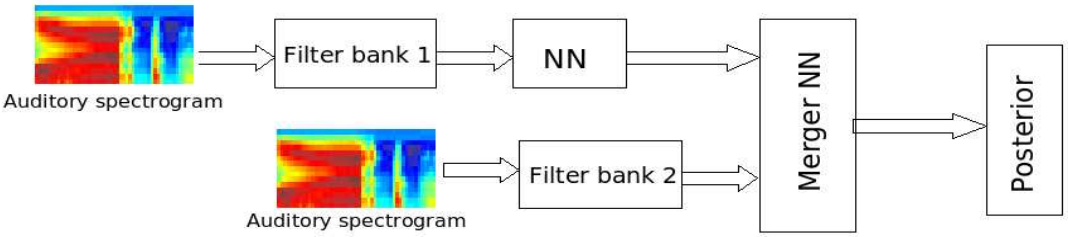

Figure 4: Hierarchical processing of modulation spectrum frequencies. Contrarily to parallel processing the order in which modulation frequencies are processed matters.

and G-low as follows:

$$\begin{aligned} \text{G-high} &= [\text{G1-high}, \text{G2-high}] = [g1_{\sigma_i}, g2_{\sigma_i}] \quad (3) \\ &\text{with} \quad \sigma_i = \{0.8\,,\ 1.2\,,\ 1.8\} \\ \text{G-low} &= [\text{G1-low}, \text{G2-low}] = [g1_{\sigma_i}, g2_{\sigma_i}] \quad (4) \\ &\text{with} \quad \sigma_i = \{2.7\,,\ 4\,,\ 6\} \end{aligned}$$

Filters G1-high and G2-high are short filters (figure 1 continuous lines) and they process high modulation frequencies (figure 2 continuous lines). Filters G1-low and G2-low are long filters (figure 1 dashed lines) and they process low modulation frequencies (figure 2 dashed lines). We present in the following experiments to asses if their combination should happen in parallel or sequential fashion.

Features	PLP	MRASTA	G-high	G-low	Comb G-high/G-low	Hier G-high to G-low	Hier G-low to G-high
WER	42.4	45.8	45.9	50.0	41.4	40.0	45.8

Table 1: Summary of RT05 WER for all experiments.

3 System description

Experiments are run with the AMI LVCSR system for meeting transcription described in [11]. 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 [12] 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 [11]. Juicer large vocabulary decoder [13] is used for recognition with a pruned trigram language model.

Table 2 reports results for the PLP plus dynamic features system and the MRASTA-TANDEM system. Both these baseline feature sets are obtained by training a single Neural Network on the

whole training set in order to obtain estimates of phoneme posteriors.

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 2: RT05 WER for Meeting data: baseline PLP system and MRASTA features

4 Parallel Processing

In the first set of experiments, a separate neural network for estimating phoneme posterior probabilities is trained for each part of the modulation spectrum. Those outputs can be combined together to provide a single phoneme posterior estimation. The process is depicted in figure 3.

In a first step the auditory spectrum is filtered with filter-banks G-high and G-low. This will provide two representations of the auditory spectrum at different time resolutions. Two independent neural networks are trained on high and low modulation frequencies; their output is recombined using a neural network merger classifier. The merger neural network takes as input 9 consecutive frames from previous neural networks. Final posterior distributions are transformed using the TANDEM scheme for use in the LVCSR system.

Table 3 shows results for high and low modulation frequencies and for combination of high/low frequencies.

Features	TOT	AMI	CMU	ICSI	NIST	VT
G-high	45.9	48.7	41.9	37.3	53.3	49.2
G-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 3: RT05 WER for high, low modulation frequencies and combination

Features obtained using filter-bank G-high have the same overall performance of full MRASTA filter-bank. However, features obtained using G-low have noticeably worse performance. The combination of high and low modulation frequencies using a merger classifier reduces WER by 4.4% w.r.t. the single classifier scheme and outperforms by 1% the PLP baseline. This experiment shows that separate processing of different modulation frequency channels is beneficial compared to using a single modulation frequency channel. The improvement is verified on all RT05 subsets.

5 Hierarchical processing

In this section, we consider hierarchical (sequential) processing of modulation frequencies. In these experiments we will use two separate modulation frequency channels as described above. The proposed system is depicted in figure 4. Critical band auditory spectrogram is processed through a first modulation filter bank followed by a NN to obtain phoneme posteriors. These posteriors are then concatenated with features obtained by processing the spectrogram with a second filter-bank. These two concatenated vectors then form an input to a second phoneme posterior-estimating NN. In such a way, phoneme estimates from the first net are modified by a second net using an evidence from a different range of modulation frequencies. This NN topology is similar to the one we used in [9].

In contrary to parallel processing , the order in which modulation frequencies are presented does make a difference. In table 4 we report WER for features obtained both moving from high to low and from low to high modulation frequencies.

Moving in the hierarchy from low frequencies to high frequencies yields similar performance as a single MRASTA neural network. On the other hand, moving from high to low modulation frequencies

Features	TOT	AMI	CMU	ICSI	NIST	VT
G-low to G-high	45.8	48.3	43.5	37.0	52.5	48.5
G-high to G-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.

produce a significant reduction of 5.8% into final WER w.r.t. single classifier approach. 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 carry out an additional experiment. Posterior features from the single MRASTA neural network that processes all frequency modulation simultaneously are presented as input to a second NN. The second NN does not use additional input but only re-processes a block of concatenated posterior features.

Features	TOT	AMI	CMU	ICSI	NIST	VT
Hier Posterior	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. Hierarchical processing improves performances w.r.t. MRASTA of 1.6% absolute. However it does not reach WER of architecture in figure 4. This means that the improvements are actually coming from the sequential processing of modulation frequencies and not from the hierarchical classifier itself.

6 Summary and Discussions

Motivated by some recent findings in physiology [14] and psychophysics [5] [8] of auditory processing, we investigated parallel and hierarchical processing of different parts of the modulation spectrum. Modulation frequency filter-bank applied in these experiments has been proposed earlier in [4] for ASR application and is referred as MRASTA. In previous related works, experiments have been conducted using a single classifier.

The current work differs in exploring multiple classifying channels and explores both parallel and hierarchical processing architectures using TANDEM approach. Table 1 summarize results of all previous experiments.

Baseline PLP system outperforms the single net MRASTA features. For the further experiments, MRASTA filter bank is separated into two set of filter banks referred as G-low and G-high. In parallel architecture (see figure 3) two independent Neural Networks are trained on G-low and G-high and their outputs combined. This approach reduces WER of 4.4% absolute w.r.t. the single Neural Network approach and outperforms baseline PLP system by 1%.

Further, we investigated the use of hierarchical processing as in figure 3 in which different modulation frequencies are processed in a hierarchical fashion. When the classification is done first on the high modulation frequency data and the output from this classifier is combined with data from lower modulation frequency range, a 5.8% improvement is obtained (this system also outperforms baseline PLP system by 2.4%), while when processing order goes from low to high frequencies, overall WER is similar to the use of MRASTA with a single NN classifier.

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 additional input from the time-frequency plane. This reduces WER by 1.6% but does not achieve recognition rates of architecture in figure 4.

To summarize, separate processing of modulation frequencies lowers considerably WER compared to approaches that uses single classifier. Out of the two proposed methods, hierarchical processing is outperforming parallel processing. Improvements are verified on all subset of the RT05 evaluation data. We found that the best performance is obtained when the classification is first done on high modulation frequencies and data from low modulation frequency range are added to phoneme posteriors from the first probability estimation step. This is in principle consistent with hierarchical processing observed in mammalian auditory system [8].

7 Acknowledgments

This work was supported by the European Community Integrated Project DIRAC IST 027787 and by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). Authors would like to thanks Dr. Jithendra Vepa, Dr. Thomas Hain and the AMI ASR team for their help with the LVCSR system.

References

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

[2] Kingsbury B.E.D., Morgan N., and Greenberg S., "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] Hermansky H. and Fousek P., "Multi-resolution rasta filtering for tandem-based asr.," in *Proceedings of Interspeech 2005*, 2005.

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

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

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

[8] Miller et al., "Spectro-temporal receptive fields in the lemniscal auditory thalamus and cortex," *The journal of Neurophysiology*, vol. 87(1), 2002.

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

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

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

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

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

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