IDIAP RESEARCH REPORT

On the Combination of Auditory and Modulation Frequency Channels for ASR applications

Fabio Valente [a] and Hynek Hermansky [a]

IDIAP–RR 08-12

June 2008

published in
Interspeech 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 08-12

On the Combination of Auditory and Modulation Frequency Channels for ASR applications

Fabio Valente and Hynek Hermansky

June 2008

published in
Interspeech 2008

Abstract. This paper investigates the combination of evidence coming from different frequency channels obtained filtering the speech signal at different auditory and modulation frequencies. In our previous work [1], we showed that combination of classifiers trained on different ranges of *modulation* frequencies is more effective if performed in sequential (hierarchical) fashion. In this work we verify that combination of classifiers trained on different ranges of *auditory* frequencies is more effective if performed in parallel fashion. Furthermore we propose an architecture based on neural networks for combining evidence coming from different auditory-modulation frequency sub-bands that takes advantages of previous findings. This reduces the final WER by 6.2% (from 45.8% to 39.6%) w.r.t the single classifier approach in a LVCSR task.

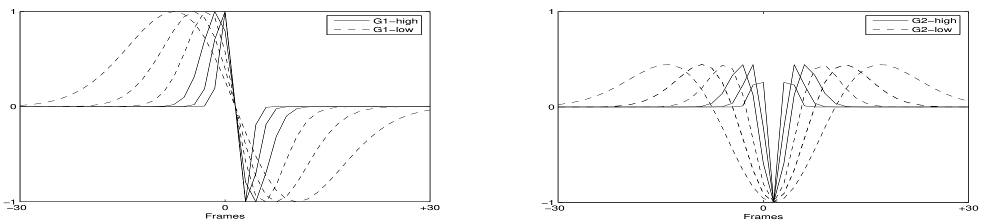

Figure 1: Set of temporal filter obtained by first (G1 left picture) and second (G2 right picture) order derivation of Gaussian function. G1 and G2 are successively split in two filter bank (G1-low and G2-low, dashed line) and (G2-high and G2-high continuous line) that filter respectively high and low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

1 Introduction

Typical Automatic Speech Recognition (ASR) features are obtained through short term spectrum of 20-30 ms segments of speech signal. This representation only extracts instantaneous frequency components of the signal. Power spectrum is then integrated using a bank of filters equally spaced on an auditory scale (e.g. Bark scale or Mel scale) thus obtaining the *auditory spectrum*.

The spectral dynamics are generally recovered from the short-term spectrum adding dynamic features (temporal differentials of the spectral trajectory at the given instant) or considering the *modulation spectrum* of the signal (long segments of spectral energy trajectories obtained by STFT [2],[3]).

Combining information obtained training separate classifiers on different frequency sub-bands of the original signal is a well accepted practise in speech recognition. Separate processing of *auditory* frequency ranges has been proposed in the multi-band framework (see [4],[5]). This method is based on the idea that the human hearing provides for performing the recognition independently in each frequency sub-band and the final decision can be taken merging decisions obtained from each sub-band. Multi-band speech recognition is particularly useful in noisy conditions.

Motivated by findings in modulation frequencies perception [6], in our previous work [1], it is shown that separate processing of *modulation* frequency sub-bands can significantly improve the performance in ASR systems. Furthermore the use of parallel and sequential combination is investigated showing that sequential processing moving from high to low modulation frequencies outperforms parallel processing.

In this paper, we first investigate whether the combination of auditory frequency sub-bands should be parallel or hierarchical completing thus the results of [1]. Then we consider combination of *joint auditory-modulation* frequency sub-bands and we investigate the relative improvements w.r.t. the single classifier approach. Those methods are motivated by physiological findings in auditory processing (see [7]) and are currently investigated by the speech processing community for application to ASR (see [8],[9]). Contrary to previous related works, we pursue the investigation in a large vocabulary task.

The paper is organized as follows: section 2 describes the preprocessing steps for extracting the auditory-modulation frequency channels, section 3 describes the experimental setting and the LVCSR system used for experiments, section 4 investigate parallel and sequential (hierarchical) combination of auditory frequencies only. Section 5 describes experiments on combination of auditory-modulation frequencies and finally in section 6 we discuss results.

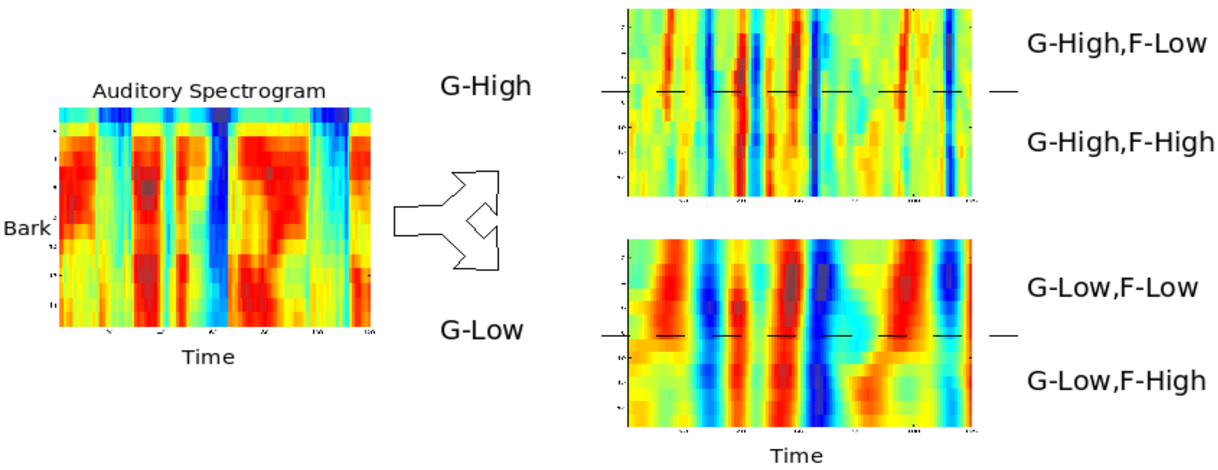

Figure 2: Auditory-modulation frequency channels extraction: auditory spectrum is filtered with a set of Gaussian filters that extracts high and low modulation frequencies (G-High and G-Low). After that auditory frequencies are split in two channels (F-High and F-Low) .This produces four different channels.

2 Time-frequency processing

This section describes the processing for extracting evidence from different auditory-modulation frequency sub-bands. 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. 15 critical bands are used.

In order to extract different modulation frequencies, MRASTA filtering is then applied (see [10] for details). 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) and second derivatives (G2) of Gaussian functions with variance σ_i varying in the range 8-130 ms (see figure 1). They have constant bandwidth on a logarithmic scale. This processing is qualitatively consistent with the perceptual model proposed in [6]. 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. As proposed in [1], filter-Banks G1 and G2 (6 filters each) are split in two separate filter-banks referred as G-Low and G-High that filter respectively high and low modulation frequencies. This produces $15 \times 6 + 15 \times 6 = 180$ features. Identical filters are used for all critical bands. Then, as in the conventional multi-band framework, the auditory frequencies are split into two sub-bands of respectively 7 and 8 critical bands. We refer to those as F-Low and F-High.

In this way the information about the signal is divided in four auditory-modulation frequency channels: (F-Low,G-Low),(F-Low,G-High),(F-High,G-Low),(F-High,G-High) that represent all combination of low and high auditory and modulation frequencies. This processing is depicted in figure 2.

For separate processing of modulation frequency sub-bands see [1]. In section 4 we analyze separate processing of auditory sub-bands and in section 5 we analyze the joint processing of auditory-modulation sub-bands.

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

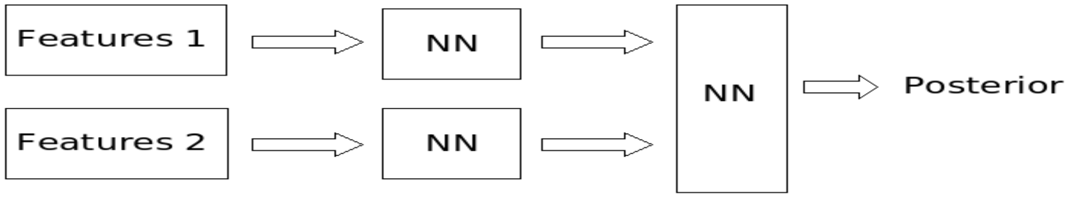

Figure 3: Parallel processing of two feature sets.

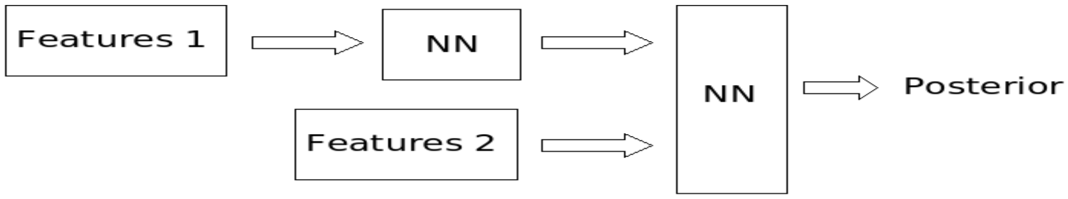

Figure 4: Hierarchical processing of two feature sets.

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 1 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. The PLP based system outperforms the MRASTA based 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 Meeting data: baseline PLP system and MRASTA features

4 Hierarchical and parallel processing of auditory frequencies

In this section, we investigate from an ASR perspective whether classifiers trained on separate auditory frequency ranges should be combined in parallel or hierarchical fashion. We limit our investigation to the two channel F-Low and F-High. In this experiment no splitting of modulation frequencies is considered.

In parallel processing, a separate neural network for estimating phoneme posterior probabilities is trained for each part of the auditory spectrum. Those outputs can be combined together to provide a single phoneme posterior estimation using a merger neural network (see figure 3). Before being used in the merger NN, posteriors are modified according to a log/KLT transform.

In hierarchical processing, a NN is trained on a first auditory frequency range to obtain phoneme posteriors. These posteriors are modified according to a log/KLT transform and then concatenated with features obtained from the second auditory frequency range thus forming 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 auditory frequencies (see figure 4). In contrary to parallel processing , the order in which frequencies are presented does make a difference.

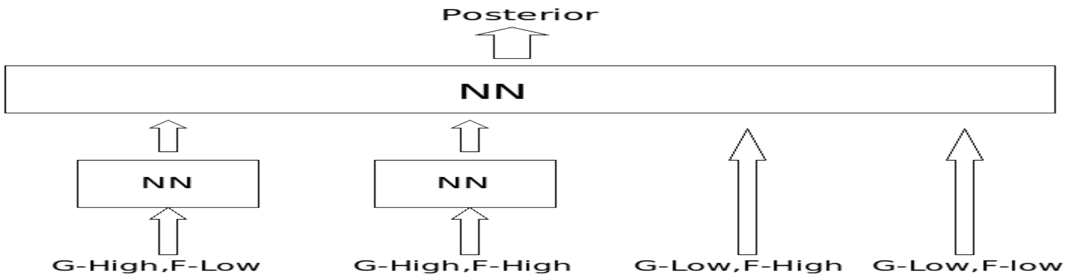

Figure 5: Proposed architecture for combining auditory and modulation frequency channels: auditory frequency channels are processed in parallel and modulation frequency channels are processed in sequence (hierarchically).

In table 2, we report WER for features obtained both moving from high to low and from low to high auditory frequencies and for features obtained processing auditory frequencies in parallel.

Features	TOT	AMI	CMU	ICSI	NIST	VT
MRASTA	45.8	47.6	41.9	37.1	53.7	49.7
F-low to F-high	45.0	48.1	42.9	36.0	51.0	47.7
F-high to F-low	44.3	**45.7**	42.5	**35.2**	50.5	48.8
Parallel	**43.9**	46.0	**41.0**	35.8	**50.7**	**47.1**

Table 2: RT05 WER for Hierarchical modulation frequencies processing: from low to high and from high to low frequencies.

In case of combination of auditory frequencies, parallel processing outperforms hierarchical processing. Furthermore using two separate frequency channels reduces the WER by 2% absolute w.r.t. the single classifier approach.

5 Parallel auditory-modulation frequency processing

In this section, we propose an experiment close in spirit to the work of [14]. A separate NN classifier is trained on each of the four sub-bands obtained as described in section 2 in order to estimate four sets of phoneme posteriors. WER for each of those individual classifiers are reported in table 3. Posteriors are then combined into a single feature stream using a NN as merger classifier. This approach combines in parallel all the four information channels of figure 2. Results are reported in table 4.

	G-Low	G-High
F-Low	65.9	66.8
F-High	**65.0**	67.3

Table 3: WER for all the four auditory-modulation frequencies channels processed separately

Separate processing of auditory-modulation frequency channels reduces the WER from 45.8% to 40.7%.

To further investigate the contribution of each channel, we report in the following WER for each of the splitting depicted in figure 2.

Let us consider two auditory frequency sub-bands for each of the two modulation frequency sub-bands. In other words output from separate classifiers trained on sub-bands (F-Low,G-Low) and (F-High,G-Low) is merged into a single feature stream M-Low and similarly output from separate classifiers trained on sub-bands (F-Low,G-High) and (F-High,G-High) is merged into a single feature

Features	TOT	AMI	CMU	ICSI	NIST	VT
	40.7	40.7	38.6	32.6	47.6	44.5

Table 4: WER for parallel combination of evidence from four auditory-modulation frequency channels.

stream M-high. This approach is equivalent to the conventional multi-band framework where the sub-bands are formed at different modulation frequencies. Results are reported in table 5.

Features	TOT	AMI	CMU	ICSI	NIST	VT
M-High	**40.7**	42.6	38.3	32.6	46.7	43.9
M-Low	46.2	48.8	44.0	37.1	52.4	49.5

Table 5: WER for multi-band approach at different modulation frequencies

Features obtained combining auditory frequencies at high modulation frequencies outperforms those obtained at low modulation frequencies. Performance of feature M-High is already similar to those obtained merging all the four channels (see table 5) thus information provided from low modulation frequencies is not effectively incorporated when parallel combination is used.

In [1], it is shown that combination of classifiers trained on separate ranges of modulation frequencies is more effective when performed in hierarchical fashion. In section 4, it is shown that combination of classifiers trained on separate ranges of auditory frequencies is more effective when performed in parallel fashion. In the previous section, we verified that parallel combination of auditory-modulation frequency channels does not take benefit of information coming from low modulation frequency channels.

Thus we propose an architecture for processing in parallel fashion auditory frequencies at high modulation frequencies and hierarchically incorporating low modulation frequencies. This architecture is depicted in figure 5.

Results for features obtained according to this method are reported in table 5.

Features	TOT	AMI	CMU	ICSI	NIST	VT
	39.6	41.9	37.0	31.9	45.1	42.9

Table 6: WER for proposed architecture of figure 5

When low modulation frequencies are introduced in hierarchical fashion WER further reduce of 1.1%. Thus processing frequency channels as in figure 5 also take benefit of information coming from low modulation frequency ranges.

6 Conclusions and discussions

In this work we address the combination of evidence coming from different auditory-modulation frequency sub-bands. This study complete our previous work on separate processing of modulation frequencies for ASR purposes. We can conclude that:

- In combination of classifiers trained on different modulation frequency ranges, sequential (hierarchical) combination outperform parallel combination when processing moves from high to low frequencies (see [1]). Those conclusions are consistent with physiological findings in [15].

- In combination of classifiers trained on different auditory frequency ranges, parallel combination outperform sequential (hierarchical) combination. Those conclusions are consistent with several previous works on speech perception [16] and multi-band speech recognition [4],[5].

We then investigate the combination of classifiers trained on four different channels that combine high and low modulation and auditory frequencies (see figure 2). This experiment is similar in spirit to the work of [14]. We verify that in a LVCSR task, WER reduces by 5% absolute (from 45.8% to 40.7%) w.r.t. the single classifier approach.

However WER for combination of four channels is actually equivalent to WER obtained combining only auditory sub-bands at high modulation frequencies. This indicates that evidence coming from low modulation frequencies is not adding any extra information.

For this reason we propose the architecture depicted in figure 5 based on the fact that modulation frequencies are combined hierarchically and auditory frequencies are combined in parallel. Separate auditory frequencies at high modulation frequencies are processed in parallel first and subsequently auditory frequencies at low modulation frequencies are introduced in hierarchical fashion. This further reduces WER by 1% w.r.t. the parallel combination of the four channels and by 6.2% w.r.t. the single classifier approach.

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] Valente F., Vepa J., and Hermansky H., "Hierarchical and parallel processing of modulation spectrum for asr applications," in *Proceedings of ICASSP*, 2008.

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

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

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

[5] Bourlard H. and Dupont S., "A new asr approach based on independent processing and recombination of partial frequency bands," in *Proceedings of ICSLP*, 1996.

[6] 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.

[7] 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.

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

[9] Mesgarani N., Stephen D., and Shamma S., "Representation of phonemes in primary auditory cortex: how the brain analyzes speech," *Proc. of ICASSP*, 2007.

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

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

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

[16] J.B. Allen, "How do humans process and recognize speech?," *IEEE Transactions on Speech and Audio Processing*, vol. 2, Oct. 1994.