IDIAP RESEARCH REPORT

Multi-resolution RASTA filtering for TANDEM-based ASR

Hynek Hermansky [a], Petr Fousek [a b]

IDIAP–RR 05-18

April 2005

submitted for publication

[a] IDIAP Research Institute, Martigny, Switzerland
[b] CTU Prague, Faculty of Electrical Engineering, Prague, Czech Republic

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 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 05-18

Multi-resolution RASTA filtering for TANDEM-based ASR

Hynek Hermansky, Petr Fousek

April 2005

submitted for publication

Abstract. New speech representation based on multiple filtering of temporal trajectories of speech energies in frequency sub-bands is proposed and tested. The technique extends earlier works on delta features and RASTA filtering by processing temporal trajectories by a bank of band-pass filters with varying resolutions. In initial tests on OGI Digits database the technique yields about 30% relative improvement in word error rate over the conventional PLP features. Since the applied filters have zero-mean impulse responses, the technique is inherently robust to linear distortions.

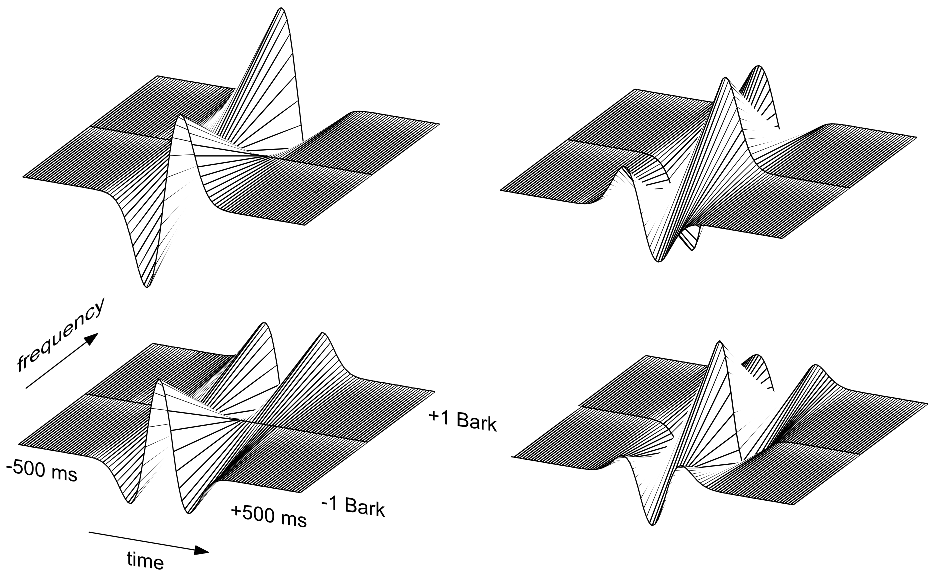

Figure 1: *Example of impulse responses of two-dimensional RASTA filters.*

1 Introduction

RASTA filtering has been proposed more than a decade ago as a means for alleviating effect of linear distortions of the signal [1]. The basic idea was similar to the original intended use of dynamic features [2], i.e. to derive feature representation that would be less sensitive to linear distortions of the signal as introduced e.g. by varying acoustic and communications channels. In the original RASTA concept, temporal trajectories of critical-band logarithmic spectral energies were filtered by a band-pass filter, which passed only these modulation frequencies that carry most of message-specific information. The filter architecture and some of its parameters were designed *ad hoc*, only a single parameter (the single pole of the filter) was found by minimizing word recognition error in a simple recognition experiment. The filter characteristics turned out to be quite consistent with several hearing phenomena [3].

The current work extends the RASTA approach by applying a bank of two-dimensional band-pass filters as a pre-processing step in TANDEM feature extraction. The filters are formed by temporal differentiation and double-differentiation and frequency differentiation of temporal trajectories of critical-band spectral energies smoothed by a Gaussian function with varying width.

The paper is organized as follows. The second section reviews several past approaches that are related and that lead to the current work. Next we describe the current technique. Experimental results in recognition of connected digits and in recognition of selected words from a large vocabulary continuous speech DARPA task are given. We also demonstrate potential for robustness of the technique to linear distortions of the test data.

2 Past related works

RASTA filter-banks: Later efforts in designing a set of RASTA filters using linear discriminant analysis (LDA) technique yielded close-to-zero-phase filters with frequency responses similar to response of the original RASTA filter and its temporal derivatives [4]. Kanedera and his colleagues [5] experimented with filtering temporal trajectories of speech features through a set of multiple filters with impulse responses representing harmonic functions with increasing spectral resolution, covering the modulation frequency range found important for perception of speech [6]. Similar experiments were later carried out by Tyagi et al [7]. Hermansky and Sharma [8] proposed a large set of RASTA

filters, each representing a matched filter for an averaged coarticulation pattern of a given phoneme at a given frequency (mean TRAP approach).
2-D RASTA filters: Jain and Hermansky [9] derived two-dimensional time-frequency RASTA filters using PCA analysis that were used in conjunction with MLP-based TRAP classification. The filters resembled cosine functions in temporal domain and frequency averaging and first and second differences in frequency domain. The optimal frequency-domain filters span was around 3 critical bands. In a related work, Grezl and Hermansky [10] demonstrated advantage of use of explicit cosine functions together with frequency differentiation.
Posteriors as features for conventional ASR (TANDEM): To deal with a high-dimensional representation of the original speech signal (the number of local features in TRAP based classification is given by a product of number of frequency-specific classes with a number of frequency bands, resulting in as many as several hundred features produced at the 100 Hz sampling rate), they used a multi-layer perceptron feed-forward neural net (MLP) to convert this large feature space into posterior probabilities of phonemes. Hermansky et al. [11] also proposed a way of converting these posteriors to features appropriate for a conventional HMM recognizers. This hierarchical classification technique of combining various information sources in deriving features for a conventional HMM-based recognizer came to be known as TANDEM feature extraction. Another hierarchical classification approaches, similar in spirit to TANDEM but differing in details, are layered HMMs [12], gamma-features [13], and FMPE features [14].
Emulating auditory cortical receptive fields: TANDEM approach in conjunction with multiple 2-D filtering of time-frequency plane was applied in differently motivated but related efforts using two-dimensional time-frequency Gabor filters [15] who explicitly stated relation of such speech processing with known physiology of auditory cortex. They attempted for a simplified version of Shamma's model of cortical processing [16],[17]. Our current work is most closely related both in motivation and in spirit to the work of Kleinschmidt and Gelbart.

3 Multi-resolution features

Combination of temporal and frequency filters applied to critical-band spectrogram can be interpreted as a 2-D filtering of the spectro-temporal plane. In our experiments we implemented the 2-D filtering by first processing critical band trajectories with temporal filters and subsequently applied frequency filters to the result, see diagram at Fig. 2.

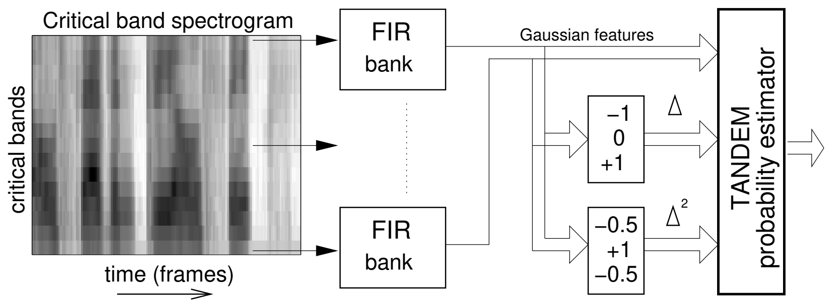

Figure 2: *Feature extraction scheme.*

Critical-band auditory spectrum is extracted from a signal every 10 ms as it is common to PLP technique [18]. By filtering temporal trajectory of each critical band with a bank of N fixed-length low-pass FIR filters representing Gaussian functions of several different widths and by subsequent computing first and second differentials of the smoothed trajectories we would obtain a set of $N \times 2$ modified spectra at every frame (Gaussian features). The same filter-bank is used for all bands.

3.1 Temporal filters

Instead of filtering sub-band trajectories with the Gaussian function and subsequently computing the differentials, we use directly the discrete versions of the first and second analytic derivatives of a Gaussian function as impulse responses. Then the filter impulses are given by

$$g_1[x] \propto -\frac{x}{\sigma^2}\ \exp(-\frac{x^2}{2\sigma^2}), \tag{1}$$

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

where x is time, $x \in \langle -500, 500 \rangle$ ms with the step of 10 ms; standard deviation σ determines the effective width of the Gaussian. 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 filters, i.e. they are centered 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. Impulse responses given by Eq. (1) are shown in the left part of Fig.3, the right parts shows impulse responses given by Eq. (2). Respective frequency responses are illustrated in Fig. 2.

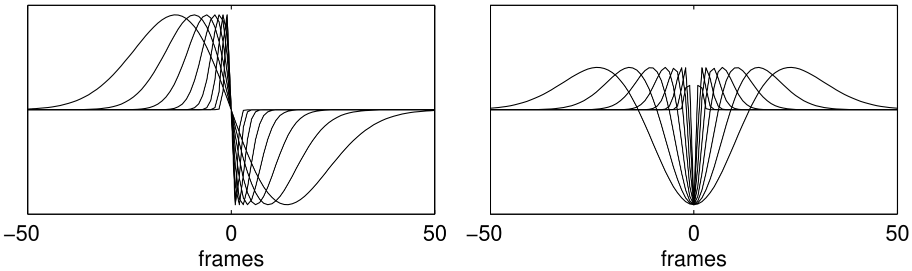

Figure 3: *Normalized impulse responses of the two sampled and truncated Gaussian derivatives for σ = 8 – 130 ms.*

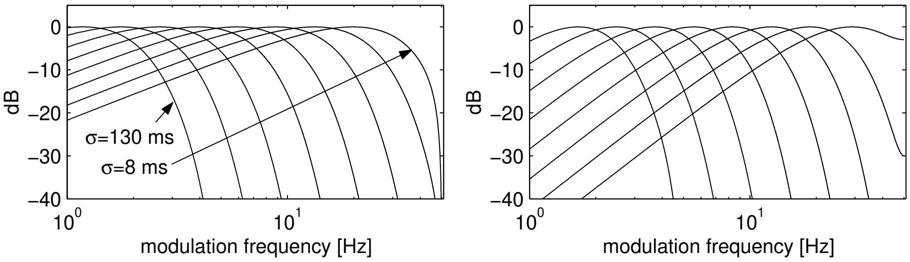

Figure 4: *Normalized frequency responses of first two sampled and truncated Gaussian derivatives for σ = 8 – 130 ms.*

Since we use discrete impulse responses of the length 101 samples, we approximate real Gaussian derivatives with certain error, which increases towards both extremes of the σ value. Using somehow arbitrary criterion that DC offset of the sampled impulse response must not excess 10% of the maximal absolute value of the response, we get a range $\sigma \in (6, 130)$ ms. For smaller σ values the sampling is too sparse, for larger σ values there are significant discontinuities at the endpoints due to the finite truncation of the infinite Gaussian function, both introducing DC offset [1]. In our experiments we use logarithmically spaced impulse responses in a σ range 8–130 ms, see Fig. 3.

3.2 2-D filters

Subsequently, first and second frequency derivatives are approximated by 3-tap FIR filters with impulse responses $\{-1, 0, +1\}$ and $\{-0.5, 1, -0.5\}$, introducing three-Bark frequency context. Feature vector, which forms an input to the MLP in TANDEM probability estimator, is obtained by concatenation of modified auditory spectra. Examples of impulse responses of 2-D filters with $\sigma = 60$ ms are shown in Fig. 1.

4 Experiments

This paper reports experiments with three feature streams:

- Gaussian features,
- Gaussian features + 1^{st} frequency derivative,
- Gaussian features + 1^{st} frequency derivative + 2^{nd} frequency derivative.

We report on experiments on a small-vocabulary continuous digit recognition (OGI Digits database). Since the back-end for proposed features was based on a TANDEM approach, our baselines were PLP-TANDEM and TRAP-TANDEM classifiers. In former case, the MLP was trained to estimate posterior probabilities of 29 English phonemes using the whole Stories database plus the training part of Numbers95 database. In case of TRAP-TANDEM, band-specific MLPs were trained on Stories and a merger MLP on the training part of Numbers95. Approximately 10% of data were used for cross-validation. Posteriors estimated by TANDEM MLPs were passed through logarithm and PCA transform with bases derived from training data. Final Gauss-TANDEM features were fed to a phoneme-based GMM/HMM system with 22 context-independent three-state phoneme HMMs, each model distribution represented by 32 Gaussian mixture components[2]. HMMs were trained on training part of Numbers95. Recognized were eleven (0-9 and "zero") digits in 28 pronunciation variants. Results are shown on frame error rate (FER) and on word error rate (WER). FER was defined on MLP posteriors as a ratio of frames with maximum posterior matching the underlying class label over the overall number of frames, considering MLP cross-validation set. WER was defined as WER=(S+I+D)/N, where S, I, D, N are counts of substitutions, insertions, deletions, and all recognized words, respectively. Results for the baselines in per cent are shown in tab. 1.

system	FER[%]	WER[%]
PLP	–	5.2
PLP-TANDEM	18	4.8
TRAP-TANDEM	19	4.7

Table 1: Performance of baseline systems in terms of frame error rate (FER) and word error rate (WER).

[1] Actually, the first sampled derivative of Gaussian function has odd symmetry and has always zero mean, but the second sampled and/or truncated derivative may have non-zero mean.

[2] Only HMMs for phonemes contained in ten recognized digits, were trained. For the EM training, all files containing other phonemes than those 22 were removed from the training set.

4.1 Extracting 656 features

First, the bank of 16 filters, consisting of first and second order derivatives of Gaussian functions as described in Section 3.1 was applied to all 15 temporal trajectories of critical-band spectral energies at all frequencies, yielding $16 \times 15 = 240$ features per frame. These formed the main feature stream.

To form the other two feature streams, we emulated first and second order frequency derivatives of the stream by applying two FIR filters to outputs of each of the 16 filters, across frequencies, as described in section 3.2. Derivatives for the first and last critical bands are not defined, so we ended up with two additional feature sets, each of size $16 \times 13 = 208$ features.

The second stream was then formed by appending the first order frequency derivatives to the main feature stream. This yielded 240+208=448 features (Gauss+Δf stream). The third stream was formed by appending second order derivatives to the second stream features, yielding 448+208=656 features (Gauss+Δf+$\Delta^2 f$ stream).

All three feature streams were fed independently to an MLP, giving posteriors estimates, which were passed to the GMM/HMM back-end. Topology of the MLP differed among the three streams only in the input layer size. Performance of all systems can be seen in tab. 2.

system	FER [%]	WER [%]
Gauss-TANDEM	18	4.2
Gauss+Δf-TANDEM	17	3.6
Gauss+Δf+$\Delta^2 f$-TANDEM	18	3.7

Table 2: Performance of three proposed systems in terms of frame error rate (FER) and word error rate (WER).

All proposed feature streams outperformed the baseline. Augmenting Gaussian features with Δf features brought significant improvement, but adding $\Delta^2 f$ features slightly decreased the performance. The performance of standalone Δf features was 5.0% WER.

4.2 Effect of range of filter pass-bands

As shown in Fig. 4, the applied RASTA filters cover the whole range of modulation frequencies of the speech signal, i.e. between 1 and 50 Hz (the sampling of spectra is at 100 Hz). To get some insight into relative importance of various modulation frequencies for ASR of continuous digits, we have shrunken this range from both the lower and the higher ends. To keep the number of free parameters in the system constant, the shrinking was associated with reducing spacing between filter's center frequencies so that even for the narrowest range there were still 8 filters per frequency band. In the Fig. 5, results are plotted as a function of the modulation frequency at which the most extreme filter in the filter bank has 3 dB attenuation. WER results are shown in left part of the figure, FER results in the right part.

As shown, eliminating significant part (up to 4 Hz) of the low modulation frequency range has no noticeable effect of WER, while only moderate cut in high modulation frequencies (down to 19 Hz) is detrimental. FER shows different picture. Low modulation frequencies are more important while higher modulation frequencies can be eliminated with only minor effect of FER. A range of approximately 1.5–8 Hz appears to contain the most of the relevant information for the frame-level classification.

4.3 Channel noise robustness

To emulate a stationary channel mismatch between training and testing data, we applied first order preemphasis filter with $\alpha = 0.97$ to the test data. Such distorted test data were passed through existing systems and word error rate was evaluated. As seen in Fig. 3, while PLP and PLP-TANDEM features are very sensitive to these distortions, proposed features are quite resistant. TRAP-TANDEM

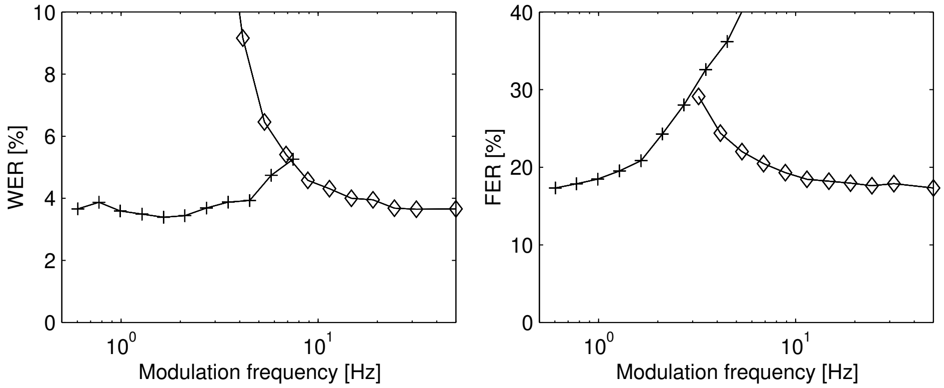

Figure 5: *Error rates as a function of 3 dB attenuation point of the most extreme filter in the filter bank.*

was also rather robust thanks to used mean normalization over TRAP length (101 samples of temporal trajectory).

system	WER[%]	relative loss [%]
PLP	13.5	160
PLP-TANDEM	12.4	160
TRAP-TANDEM	4.8	3.4
Gauss-TANDEM	4.2	2.4
Gauss+Δf-TANDEM	3.8	3.7

Table 3: Performance in channel mismatched conditions in terms of word error rate and relative performance loss.

5 Summary and Conclusions

Supported by initial results on small-vocabulary ASR experiments reported in this paper, 2-D multi-resolution RASTA filtering [5, 7, 15, 9, 10] in conjunction with TANDEM feature extraction [11] appears to be an efficient means for representing message-specific information in the speech signal. It yields considerable improvements in error rate comparing to conventional ASR features. Zero-mean impulse responses of the applied filters imply robustness to linear distortions of the signal and to changes in spectral tilt that could be induced by extra-linguistic factors, thus inherently alleviating one of major sources of harmful variability in the speech signal. Even though not extensively discussed in the present paper, the technique is consistent with the currently evolving knowledge of mammalian auditory cortical processing [16, 17].

Acknowledgments

The work was supported by DARPA under the EARS Novel Approaches grant no. MDA972-02-1-0024. The other sources of support were the IM2 Swiss National Center for Competence in Research, managed by Swiss National Science Foundation on behalf of Swiss authorities, and the European Community AMI and M4 grants.

References

[1] Hermansky, H., Morgan, N., "RASTA processing of speech", IEEE Trans. on Speech and Audio Processing, vol. 2, num. 4, pp. 578-589, Oct, 1984.

[2] Furui, S., "Speaker independent isolated word recognition using dynamic features of speech spectrum", IEEE Trans. ASSP-34, pp. 52–59, 1986.

[3] Hermansky, H., "Should recognizers have ears?", Speech Communication, 25:3-27, 1998.

[4] van Vuuren, S., Hermansky, H., "Data-Driven Design of RASTA-Like Filters", Proc. of Eurospeech 97, Rhodes, Greece, 1997.

[5] Kanedera, N. et al., "On the relative importance of various components of the modulation spectrum for automatic speech recognition", Speech Communication, 28:43–55, 1999.

[6] Arai, T. et al., "Syllable intelligibility for temporally-filtered LPC cepstral trajectories", J. Acoust. Soc. Am., 1999.

[7] Tyagi, V. et al., "Mel-Cepstrum Modulation Spectrum (MCMS) Features for Robust ASR", in IEEE ASRU, 2003.

[8] Hermansky, H., Sharma, S., "TRAPS - Classifiers of Temporal Patterns", Proc. of ICSLP'98, Sydney, Australia, 1998.

[9] Jain, P., Hermansky, H., "Beyond a Single Critical-Band in TRAP Based ASR", Proc. of Eurospeech 2003, pp. 437-440, Septemeber 2003, Geneva, 2003.

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

[11] Hermansky, H. et al., "Connectionist Feature Extraction for Conventional HMM Systems", Proc. of ICASSP 00, Istanbul, Turkey, 2000.

[12] Olivier, N. et al., "Layered representations for human activity recognition", Proc. Int. Conference on Multimodal Interfaces, pp. 3-8, 2002.

[13] Bourlard, H., et al., "Towards Using Hierarchical Posteriors for Flexible Automatic Speech Recognition Systems", Proc. of the DARPA EARS (Effective, Affordable, Reusable, Speech-to-text) Rich Transcription (RT'04) Workshop, 7-10 November 2004, IBM Palisades, NY.

[14] Powey, D. et al., "FMPE: Discriminatively trained features for speech recognition", Proc. of ICASSP 2005, pp. I-961-964, Philadelphia, 2005.

[15] Kleinschmidt, M., Gelbart, D., "Improving Word Accuracy with Gabor Feature Extraction", Proc. of ICSLP'02, Denver, Colorado, USA, 2002.

[16] Carlyon, R., Shamma, S., "An account of monaural phase sensitivity", J. Acoust. Soc. Am., 114:333-348, 2003.

[17] Elhilali, M. et al., "Intelligibility and the spectrotemporal representation of speech in the auditory cortex", Speech Communication, 41:331–348, 2003.

[18] Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech", J. Acoust. Soc. Am., Vol. 87, No. 4, April 1990.