Hierarchical Spectro-Temporal Features for Robust Speech Recognition

Xavier Domont, Martin Heckmann, Frank Joublin, Christian Goerick

2008

Preprint:

This is an accepted article published in Proc. of International Conference on Acoustics, Speech, and Signal Processing (ICASSP). The final authenticated version is available online at: https://doi.org/[DOI not available]

HIERARCHICAL SPECTRO-TEMPORAL FEATURES FOR ROBUST SPEECH RECOGNITION

Xavier Domont[1,2], *Martin Heckmann*[1], *Frank Joublin*[1], *Christian Goerick*[1]

[1]Honda Research Institute Europe GmbH
Offenbach am Main, Germany
{*firstname.lastname*}*@honda-ri.de*

[2]Technische Universität Darmstadt, Germany
Control Theory and Robotics Lab
xavier.domont@rtr.tu-darmstadt.de

ABSTRACT

Previously we presented an auditory-inspired feed-forward architecture which achieves good performance in noisy conditions on a segmented word recognition task. In this paper we propose to use a modified version of this hierarchical model to generate features for standard Hidden Markov Models. To obtain these features we firstly compute the spectrograms using a Gammatone filterbank. A filtering over the channels permits to enhance the formant frequencies which are afterwards detected using Gabor-like receptive fields. Then the responses of the receptive fields are combined to complex features which span the whole frequency range and extend over three different time windows. The features have been evaluated on a single digit recognition task. The results show that their combination with MFCCs or RASTA features yields improved recognition scores in noise.

Index Terms— Speech recognition, Robust features

1. INTRODUCTION

State of the art automatic speech recognition (ASR) systems using Mel Frequency Cepstral Coefficients (MFCCs) achieve high recognition performance on clean signals. For many applications, humanoid robotics in particular, the signals are highly noisy and the performance of conventional ASR systems decreases drastically. On the other hand, human speech perception is far less susceptible to such distortions [1]. Therefore, we believe that a higher robustness in speech recognition can be obtained by using auditory-inspired features.

Moreover, Shamma showed that the primary auditory cortex of young ferrets has a spectro-temporal organization, i.e. the receptive fields are selective to modulations in the time-frequency domain and, as in the visual cortex, have Gabor-like shapes [2]. These receptive fields have been modeled by Chin [3] and used for source separation [4] and speech detection [5]. Gabor features extraction has also been used for ASR by Kleinschmidt in [6].

In a previous work [7], we proposed a feed-forward neural network for isolated monosyllabic word recognition. This system, inspired from the visual object recognition system described in [8], contains three hierarchically-organized layers: the first layer detects local spectro-temporal patterns in whole words' spectrograms, the second one combines these patterns to more complex ones, and, finally, linear discriminant classifiers are used to build the word models. Despite good robustness against additive noise, these models required that all the words were previously segmented and that they all had the same length. A linear interpolation was performed on the spectrograms to normalize their lengths, but the performance in clean conditions was not satisfactory.

The work presented in this paper retains the principle that a spectro-temporal, hierarchical processing of the speech improves the robustness of speech recognition, but uses Hidden Markov Models (HMMs) instead of linear discriminant classifiers. These *Hierarchical Spectro-Temporal* (HIST) features permit to overcome the segmentation and normalization issue of our previous system.

In section 2, the computation and enhancement of the spectrograms are described. The hierarchical processing extracting HIST features from the spectrograms is explained in section 3. Finally, the performance on a single digits recognition task is shown in section 4. The feature extraction process is visualized in Fig. 1.

2. PREPROCESSING

The spectrograms of the speech signals are computed using a Gammatone filterbank. We used an IIR implementation of the Gammatone filterbank [9] having 128 channels ranging from 80 Hz to 8 kHz using a sampling rate of 16 kHz. The spectrograms are obtained by rectification and low-pass filtering of the filterbank response. The sampling rate of the spectrograms is then reduced to 400 Hz.

Subsequently, we perform some preprocessing on the spectrograms aiming at enhancing the formant frequency. First the influence of the speech excitation signal is compensated by emphasizing the high high frequencies by $+6$ dB per octave resulting in flattened spectrograms. Next, a set of Mexican Hat filters along the frequency axis is used to remove the harmonic structure of the spectrograms, resulting in an enhancement of the formant frequencies. For this filtering the size of the filters' kernel is channel-dependent, varying from 90 Hz for low frequencies to 120 Hz for high frequencies. This takes the logarithmic arrangement of the center frequencies in the Gammatone filterbank into account.

Figure 2 shows the enhanced spectrograms of the digit "one"

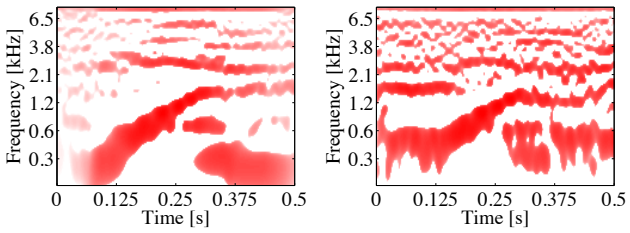

Fig. 2. Enhanced spectrogram of the digit "one" spoken by a male speaker. Without noise (a) and with babble noise added at an SNR of 5 dB (b).

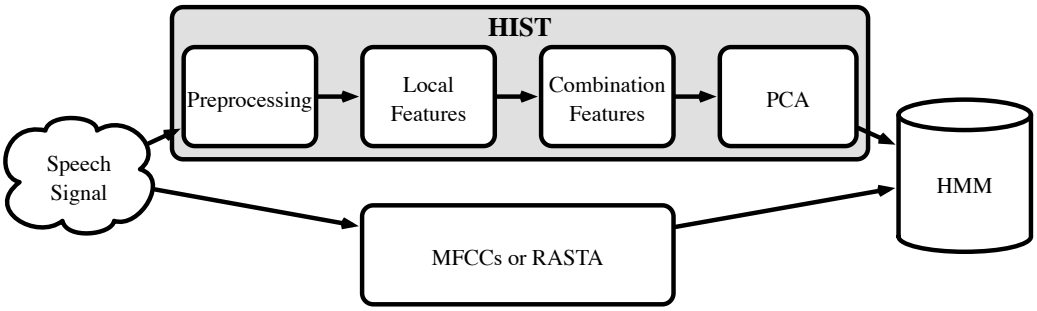

Fig. 1. Overview of the feature extraction process.

spoken by a male American speaker in clean conditions and when the signal is distorted by additive babble noise at 5 dB SNR.

3. HIERARCHICAL SPECTRO-TEMPORAL FEATURES

3.1. First stage: Extraction of local features

We want to detect local patterns in the spectrogram $\mathbf{S}$ obtained after the preprocessing described in section 2. Note that the spectrogram $\mathbf{S}$ can be interpreted as a 2D image.

The feature extraction is done by a 2D filtering with a set of n_1 receptive fields $\mathbf{w}_1^l$, taking the absolute value of the response:

$$q_1^l(t,f) = \left| \left(\mathbf{S} * \mathbf{w}_1^l \right)(t,f) \right|, \tag{1}$$

where the responses $\mathbf{q}_1^l$ of each neuron have the same size as the input spectrogram.

As in [7], $n_1 = 8$ relevant receptive fields have been learned using Independent Component Analysis (ICA) on 3500 randomly selected local 8×8 patches of the enhanced spectrograms taken from the training part of the database.

For a given point (t,f) in the spectrogram, the activity $q_1^l(t,f)$ of the lth neuron reveals how close a local patch of $\mathbf{S}$ centered in (t,f) is to the pattern l. For each local patch only the highest correlated patterns are of interest. Therefore, we perform a Winner-Take-Most (WTM) competition which inhibits the response of the less active neurons at the position (t,f):

$$r_1^l(t,f) = \begin{cases} 0 & \text{if } \frac{q_1^l(t,f)}{M(t,f)} < \gamma_1 \text{ or } M(t,f) = 0 \\ \frac{q_1^l(t,f) - \gamma_1 M(t,f)}{1-\gamma_1} & \text{else,} \end{cases} \tag{2}$$

where $M(t,f) = \max_k q_1^k(t,f)$ is the maximal value at position (t,f) over the eight neurons and $0 \leq \gamma_1 \leq 1$ is a parameter controlling the strength of the competition [8].

Furthermore, a nonlinear transformation including a threshold θ_1 is applied on all the $r_1^l(t,f)$:

$$s_1^l(t,f) = H(r_1^l(t,f) - \theta_1), \tag{3}$$

where $H(x)$ is the Heaviside step function.

Finally, the resolution of the images $\mathbf{s}_l(t,f)$ is four times reduced in both frequency and time directions, i.e. there are now 32 frequency channels and the sampling rate is 100 Hz. The images are smoothed with a 2D Gauss filter $\mathbf{g}_1$ prior to downsampling:

$$c_1^l(t,f) = \left(\mathbf{s}_1^l * \mathbf{g}_1 \right)(4t, 4f). \tag{4}$$

3.2. Second stage: Extraction of combination features

Each of the n_2 combination patterns is composed of n_1 receptive fields $\mathbf{w}_{2,l}^k$, i.e. one for each of the neurons in the previous stage. The coefficients of these receptive fields are non negative and span all frequency channels. Similarly to (1) the activity $q_2^k(t)$ of the kth neuron at the time t is given by:

$$q_2^k(t) = \sum_{l=1}^{n_1} \left(\mathbf{c}_1^l * \mathbf{w}_{2,l}^k \right)(t,f). \tag{5}$$

As the combination patterns span the whole frequency range the response of the neurons does not depend on f anymore. This means that, by computing the convolution, the patterns $\mathbf{w}_{2,l}^k$ are only shifted in the time direction. It should also be noted that the absolute value is not required in (5) as both the $\mathbf{c}_1^l$ and the $\mathbf{w}_{2,l}^k$ are non-negative.

The combination patterns are learned in an unsupervised manner using Non-Negative Sparse Coding (NNSC) [10]. NNSC differs from NMF by the presence, in the cost function (6), of a sparsity enforcing term which aims at limiting the number of non-zero coefficients required for the reconstruction. Consequently, if a feature appears often in the data, it will be learned, even if it can be obtained by a combination of two or more other features. Therefore, the NNSC is expected to learn complex and global features appearing in the data.

From the training database we compute the $\mathbf{c}_1^l$ spectrograms for the signals containing only one digit. We then cut out patches of length Δ out of these images. From these patches we learn $n_2 = 50$ combination features by minimizing the following cost function [8]:

$$E = \sum_p \|\mathbf{P}^p - \sum_{k=1}^{n_2} \alpha_k^p \mathbf{w}_2^k\|^2 + \beta \sum_p \sum_{k=1}^{n_2} |\alpha_k^p| \,. \tag{6}$$

where $\mathbf{P}^p$ is a tensor representing the n_1 layers of the pth patch, the $\mathbf{w}_2^k$ are n_2 non-negative tensors each of them containing the n_1 receptive fields $\mathbf{w}_{2,l}^k$, the α_k^p are nonnegative reconstruction factors, and β is a parameter allowing to control the sparsity of the learned features.

Three different sets of combination features have been learned for $\Delta = 40$, 80, and 160 milliseconds. For each set we obtain $n_2 = 50$ features $\left(q_2^1(t), \ldots, q_2^{n_2}(t)\right)^T$. The feature rate is 100 Hz.

For each feature set delta (resp. double-delta) features are computed using a 9th order FIR lowpass (resp. bandpass). The dimensionality of the feature vectors is then reduced from 150 to 39 using Principal Component Analysis (PCA). For each set the coefficients of the PCA are learned on the clean part of the database.

Finally the 3 feature vectors corresponding to the 3 different time windows are concatenated and the dimension is reduced to 39 using a new PCA. For computational reasons the PCA is split in two steps: firstly it is calculated on each feature set and then on the concatenated features.

4. RECOGNITION PERFORMANCE

4.1. The recognition task

Even if an evaluation on the Aurora-2 database [11] would have been desirable, due to the time intensive processing of the current Matlab implementation, we restricted this test to the single digits part of the TIDigits corpus [12]. The utterances of the test database have been mixed with additive noise in a similar way as in the Aurora-2 framework. Some differences to the Aurora-2 database have to be pointed out:

- Signals are downsampled to 16 kHz instead of 8 kHz.
- Only signals containing one single digit are kept.
- When mixing the signals with noise using FaNT [13] the G.712 is only used for the noise and signal level estimation, i.e. the obtained signals have no channel distortions.
- Three types of noise from the Noisex database [14] have been used: Babble, Factory, and Car.

Both the training and test databases contain 326 utterances for each of the 11 digits. These digits are spoken by different speakers (boys, girls, women, and men) each speaker uttering each digit twice. The speakers in the test database are different from those in the training database.

The Hidden Markov Models are trained on clean signals with HTK [15] using the same parameters as in the Aurora-2 framework [11]. Whole word HMMs contain 16 states without skip over states and a mixture of 3 Gaussians with a diagonal covariance matrix per state. See [11] for a complete description of Aurora-2's backend, the only difference with our backend being the absence of a model for pauses between words, which is irrelevant for a single digit recognition task.

4.2. Comparison with State of the Art features

In order to compare the performances of the proposed features, the recognition task has also been performed using MFCCs and RASTA-PLP features. 12 MFCCs are used, without the zeroth coefficient, with the logarithmic frame energy plus the corresponding delta and double-delta coefficients. Cepstral Mean Subtraction has been applied on the MFCCs. For the RASTA-PLP features we use an order of 14 for the linear prediction and also use delta and double-delta coefficients. In all cases the HMMs are trained on clean signals. Furthermore, the different types of features have been combined to study their complementarity.

When the signals are mixed with factory noise (Fig. 3 (a)), the performance of MFCCs decreases rapidly when the SNR is smaller than 15 dB. The RASTA features show a better robustness and perform better than MFCCs except for clean. The performance of the proposed HIST features decreases more smoothly when the noise level increases but they do not perform well on high SNRs and therefore only catch up on the RASTA features when the SNR is bellow 0 dB. The relatively poor performance of the HIST features for high SNRs can be explained by the large time windows used to compute them. However, the best word error rates (except for clean and -5 dB) are obtained by concatenating the HIST features with RASTA features. The success of this combination suggests that the proposed features use information complementary to RASTA-PLP. In clean conditions the MFCCs perform best (0.17% WER, 6 errors) but the difference with RASTA (0.2%, 7 errors) and the combination of RASTA and HIST features (0.25%, 8 errors) is not statistically significant.

In the presence of babble noise (Fig. 3 (b)), the MFCCs perform better than on factory noise but are as before outperformed by RASTA features when the SNR decreases (below 15 dB). Concerning the proposed features the two facts observed on factory noise are confirmed: the performance of the features are poor for low noise levels but the features are complementary to RASTA and their combination yields good performance in all cases. Similar results have also been observed in the presence of car noise. However this noise interferes only mildly with speech and all the features performed very well.

Figures 3 (c) and (d) show the relative improvement of the different combinations of features w.r.t. RASTA features in the presence of the two types of noise previously studied. Additionally to the combination of HIST and RASTA features, the performance of the combination of MFCCs and RASTA and a combination of MFCCs and HIST features are drawn.

With factory noise (Fig. 3 (c)), only the combination of RASTA and HIST features shows better performance than the baseline. The two other combinations perform most of the time worse than pure RASTA features. In the presence of babble noise (Fig. 3 (d)), the combination of MFCCs and RASTA features is interesting at high SNRs with, in particular, an improvement of more than 40% over RASTA features at 20 dB SNR. However, the combination of HIST and RASTA features shows, as before, the best robustness.

The better robustness of the proposed HIST features might only be due to the larger length of the analysis windows used to compute the features. In order to rule out this hypothesis, in a last set of experiments, we also used MFCCs with larger time windows. Similar to HIST features we calculated MFCCs for 40, 80, and 160 ms, with 39 coefficients. We then combined the features of these three time windows yielding 113 coefficients and applied a PCA retaining only 39 coefficients. These 39 coefficients were then combined with the standard MFCCs evaluated on 25 ms. Due to technical issues we did not use RASTA features. The results in Fig. 3 (e) and (f) show that the gain in performance using the standard MFCCs and the HIST features is larger than the one obtained by the combination of MFCCs of length 25, 40, 80, and 160 ms. Therefore, we assert that the performance improvement given by the proposed HIST features is not just due to the use of larger analysis windows but that these features also extract information which is complementary to the one captured by the conventional features.

Finally, we also applied a MVA post-processing [16] on MFCCs and HIST features. It significantly improved the performance of both type of features at SNRs below 10 dB but did not change the ranking.

5. DISCUSSION & SUMMARY

In this paper a new type of features for speech recognition has been proposed. These features are inspired from recent research on the importance of spectro-temporal processing in the primary auditory cortex. The features are computed by detecting local patterns in the spectrograms and combine them into complex patterns spanning the whole frequency spectrum.

The test on a speaker-independent single digit recognition task shows that these features alone currently are only competitive to conventional features at low SNRs. However, they seem to capture complementary information to MFCCs and RASTA features as a combi-

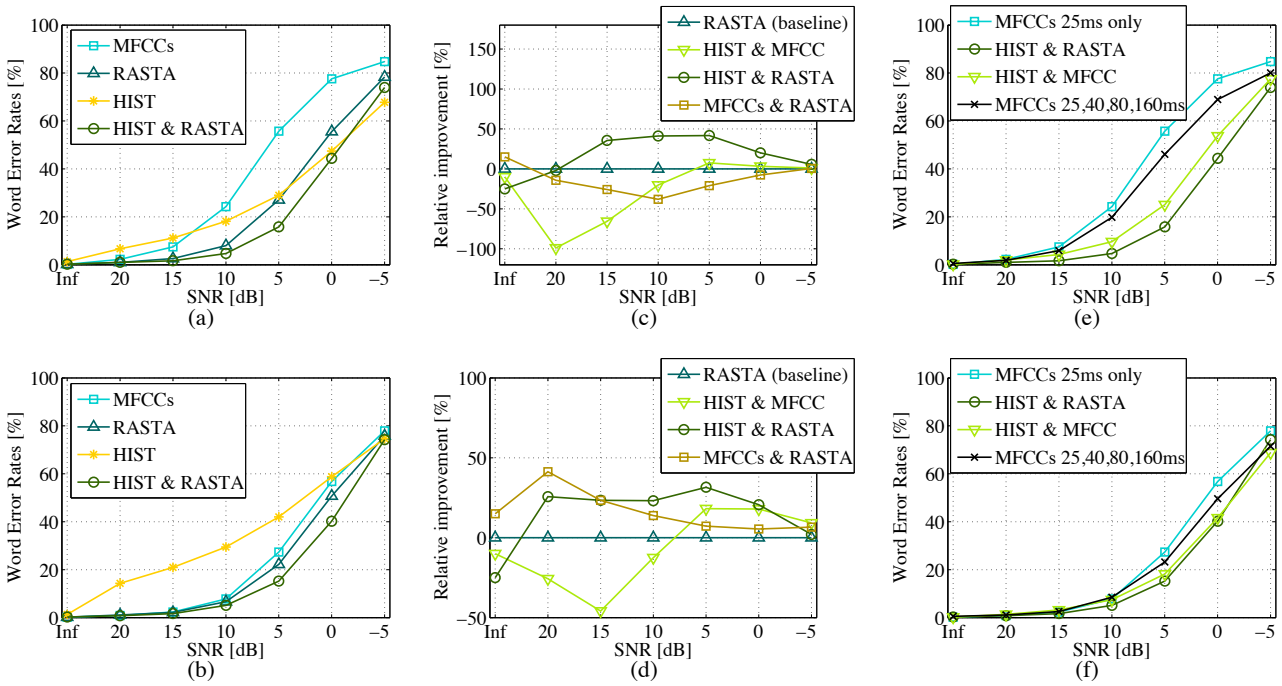

Fig. 3. Comparison of the performance in the presence of factory noise (a) or babble noise (b). Performance of the features relatively to the RASTA-PLP features in the presence of factory noise (c) or babble noise (d). Comparison of the proposed combination with MFCCs combined with MFCCs using 40, 80, and 160 ms analysis windows in the presence of factory noise (e) or babble noise (f).

nation with either one of these yields a significant higher robustness in noise. Moreover, the sampling rate used for the HIST features is rather low and the parameters of the feature extraction, and especially the competition mechanism, are not yet fully optimized, leaving substantial room for improvement of the HIST features, alone as well as in combination with the RASTA features. Tests on more challenging tasks, e.g. continuous speech recognition, are necessary to better asses the performance of the features. The single digit task is quite easy in clean condition and makes it difficult to evaluate the difference of performance between the features at high SNRs.

6. REFERENCES

[1] R.P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, 1997.

[2] S. Shamma, "On the role of space and time in auditory processing," *Trends in Cognitive Sciences*, vol. 5, no. 8, pp. 340–348, 2001.

[3] T. Chih, P. Ru, and S. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *Journal of the Acoustical Society of America*, vol. 118, pp. 887–906, 2005.

[4] M. Elhilali and S. Shamma, "A bilogically-inspired approach to the cocktail party problem," in *Proc. ICASSP*, 2006, vol. 5, pp. V–637–640.

[5] N. Mesgarani, M. Slaney, and S. Shamma, "Discrimination of speech from non-speech based on multiscale spectro-temporal modulations," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 3, pp. 920–930, 2006.

[6] M. Kleinschmidt, *Robust speech recognition based on spectrotemporal processing*, Ph.D. thesis, Uni. Oldenburg, 2002.

[7] X. Domont, M. Heckmann, H. Wersing, F. Joublin, S. Menzel, B. Sendhoff, and C. Goerick, "Word recognition with a hierarchical neural network," in *NOLISP Conference, Paris*, 2007.

[8] H. Wersing and E. Körner, "Learning optimized features for hierarchical models of invariant recognition," *Neural Computation*, vol. 15, no. 7, pp. 1559–1588, 2003.

[9] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filterbank," Tech. Rep., Apple Computer Co., 1993, Technical report #35.

[10] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[11] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR 2000, Paris, France*, 2000, pp. 181–188.

[12] R.G. Leonard, "A database for speaker independent digit recognition," in *Proc. ICASSP*, 1984, vol. 3, p. 42.11.

[13] H.-G. Hirsch, "Fant - filtering and noise adding tool," http://dnt.kr.hs-niederrhein.de/download.html.

[14] A. Varga and H.J.M. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–252, 1993.

[15] S. Young et al., "The htk book," Cambridge, December 2006.

[16] C.-P. Chen, K. Filali, and J. Bilmes, "Frontend post-processing and backend model enhancement on the aurora 2.0/3.0 databases," in *ICSLP*, 2002.