ON THE INTERPLAY BETWEEN AUDITORY-BASED FEATURES AND LOCALLY RECURRENT NEURAL NETWORKS FOR ROBUST SPEECH RECOGNITION IN NOISE

J. Tchorz[⋆], *K. Kasper*[‡], *H. Reininger*[‡], *B. Kollmeier*[⋆]

[⋆]Carl von Ossietzky-Universität, AG Medizinische Physik
26111 Oldenburg, Germany

[‡] Institut für Angewandte Physik, Johann Wolfgang-Goethe-Universität
60054 Frankfurt, Germany

tch@medi.physik.uni-oldenburg.de

ABSTRACT

The combination of a model of auditory perception (PEMO) as feature extractor and of a Locally Recurrent Neural Network (LRNN) as classifier yields promising ASR results in noise. Our study focuses on the interplay between both techniques and their ability to complement each other in the task of robust speech recognition. We performed recognition experiments with modifications of PEMO processing concerning amplitude compression and envelope modulation filtering. The results show that the distinct and sparse peaks of PEMO speech representation which are well maintained in noise are sufficient cues for LRNN-based recognition due to LRNN's ability to exploit information which is distributed over time. Enhanced envelope modulation bandpass filtering of PEMO feature vectors better reflects the average modulation spectrum of speech and further decreases the influence of noise.

1. INTRODUCTION

One major problem in automatic speech recognition (ASR) is the robustness of ASR systems against noise. Even slightly disturbed speech often leads to severe increase of the error rate, making the usefulness of the system questionable. Earlier investigations [1] have shown that a speech recognition system combining feature extraction based on a model of human auditory perception (PEMO) with a Locally Recurrent Neural Network (LRNN) as classifier is a promising approach to speech recognition in noisy environments. The combination of PEMO and LRNN yielded significantly higher isolated-word recognition rates than systems with mel-frequency cepstra or RASTA coefficients as feature vectors in combination with a discrete HMM recognizer. Adaptive J-RASTA processing [2] gave comparable results, but parts of the input signal are assumed to be speech free for noise estimation then, whereas no such assumptions are required for PEMO processing. This study focuses on a deeper investigation on the interplay between PEMO feature extraction and subsequent LRNN-based recognition. Our aim was to demonstrate the characteristics of PEMO representation of speech and to show how LRNN recognition, in contrast to HMM-based recognition, takes advantage of this kind of speech representation for robust recognition in noise. In addition, a modulation bandpass filter which reflects the average envelope modulation spectrum of speech is introduced to further decrease the influence of noise in recognition tasks.

2. RECOGNITION SYSTEM

2.1. Feature extraction

PEMO processing was originally developed to predict human performance in typical psychoacoustical temporal and spectral masking experiments [3]. The main processing steps of PEMO are (i) filtering of the digitized input signal in a basilar membrane filter bank which simulates the transfer functions of the peripheral filters, (ii) half wave rectification and low pass filtering at 1 kHz for envelope extraction in each frequency channel, (iii) an adaptive dynamic compression unit which compresses steady-state portions of the input signal almost logarithmically, whereas fast changes are transmitted linearly, and (iv) low pass filtering of the fast fluctuating envelope in each frequency channel at a cutoff frequency of 8 Hz. 17 frequency channels with center frequencies from 300 - 3300 Hz were used for feature extraction. The main characteristics of PEMO speech representation can be seen in Fig. 1. The first panel shows the filtered waveform of the German word "wiederholen" spoken by a male speaker. Plotted is one frequency channel of the filter bank corresponding to a center frequency of 720 Hz. The second panel shows the processed PEMO output of the utterance in this frequency channel. The enhanced encoding of signal onsets and offsets can be seen, as well as reduced sensitivity in an interval of recovery after the first

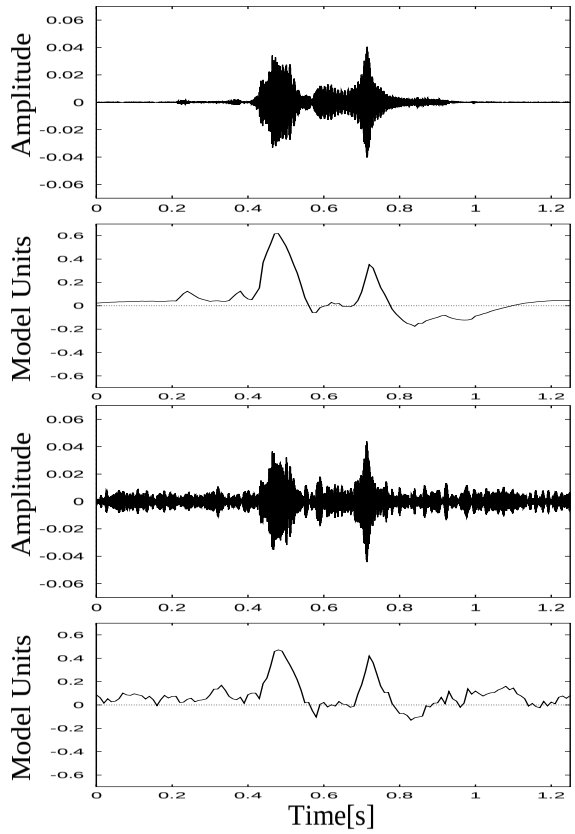

Figure 1. Example of speech representation performed by PEMO processing. See text.

peak. In the third panel, the utterance was disturbed with white noise added at 5 dB SNR before filtering. The same frequency channel as in the first panel is shown. In the last panel, the corresponding response after PEMO processing is shown. In non-speech intervals, the stationary background noise is compressed but still causes distortions in the representation compared to the undisturbed utterance. The distinct peaks, which represent changes in the input signal induced by speech portions, are not represented with the same magnitude as in the undisturbed case, but the overall structure is maintained. For further processing, the output of the auditory preprocessing is sampled at 100 Hz, the resulting feature vectors serve as input to the LRNN recognizer. PEMO processing is implemented in C-Code and takes about 1.4 times real time on an SGI RS 5000 workstation.

2.2. LRNN recognition

LRNN are biologically motivated and have been introduced in [4] in order to reduce the computational complexity of fully connected recurrent neural networks. It has been shown that ASR systems based on LRNN achieve recognition results for isolated words and connected digits which are comparable to sophisticated HMM-based systems. LRNN consists of an input layer, a hidden layer with locally recurrent connections and an output layer. The interaction between neighboring layers are unidirectional and sparse. The recurrent connections of the hidden neurons are ending at the edges of the grid. The network is trained by truncated backpropagation through time. Due to the recurrent connections in a LRNN it is possible to exploit information distributed over time in a feature sequence for classification. Compared to approaches based on Hidden Markov Models, the extraction of dynamic features is obsolete and no Viterbi algorithm for compensating varying word durations is required.

3. SPEECH MATERIAL

Recognition experiments were performed with isolated-spoken German digits. The speech material was recorded at high quality, but was filtered with a telephone transmission transfer function before feature extraction. 100 utterances of each digit from two independent sets of 100 speakers both male and female were used for training and testing. Another set of speech material was introduced to attain more realistic test conditions. It consisted of 100 utterances of each digit recorded over dialed-up telephone lines in the Berlin area.

4. EXPERIMENT I

4.1. Modifications of PEMO

The aim of the first experiment was to analyze the interplay between PEMO and LRNN. The processing step of PEMO which dominates the characteristic of the signal representation is the adaptive dynamic compression which contrasts signal on- and offsets, whereas constant portions from the input signal are suppressed. Thus, the signal representation is sparse, it contains distinct peaks rather than constant excitation over lots of time frames. To evaluate the importance of this type of signal representation for robust speech recognition with LRNN, the adaptation loops, which perform adaptive amplitude compression in PEMO processing, were replaced by a static logarithmic compression of the dynamic range (LOG). The second variation of PEMO went into the opposite direction: the emphasize of changes in the input signal was further increased, steady-state portions were compressed even more by squaring the feature values (MOD). Speaker-independent, isolated-digit recognition experiments were performed with PEMO, LOG and MOD in combination with LRNN. For comparison, recognition rates were also measured with a continuous Hidden Markov recognizer (CHMM). 5 Gaussian mixtures per state, diagonal covariance matrices and 6 emitting states per word model were used for the experiments.

	LRNN			CHMM		
	CLN	S10	TEL	CLN	S10	TEL
PEMO	98.2	89.0	93.1	93.9	46.7	69.4
MOD	97.2	89.3	92.5	92.0	54.0	64.3
LOG	98.5	10.0	10.9	92.7	26.1	66.0
FILT	98.5	95.1	94.5	93.4	50.7	80.2

Table 1. Speaker-independent recognition rates in per cent from experiments I (first three rows) and II (last row). CLN: clean speech. S10: speech simulating noise added at 10 dB SNR to the test material. TEL: real telephone speech for testing.

Speaker-independent recognition rates were measured on three sets of test data: undisturbed speech (CLN), speech which was distorted by additive speech simulating noise at 10 dB SNR before feature extraction (S10), and speech recorded via telephone lines (TEL). The recognition rates are shown in Table 1. For LRNN, no significant differences can be observed between PEMO and MOD. With LOG, high recognition rates are yielded in clean speech, but the rate drops to chance when the test material is disturbed by additive (S10) or convolutive noise (TEL). For the CHMM recognizer, the three types of features allow comparable results in clean speech. In disturbed speech, PEMO or MOD feature extraction helps to increase recognition rates significantly compared to fixed compression in LOG, but by far not to the extend as in LRNN classification. The results indicate that distinct and sparse coding of the input signal which emphasizes changes rather than constant portions (PEMO and MOD) leads to robust recognition in combination with LRNN. If the prominent peaks which encode the temporal evolution of the signal are missing (LOG), no sufficient cues for LRNN recognition are left in disturbed speech.

4.2. Manipulating the features

We analyzed the contribution of sparse and distinct peaks from PEMO processing to robust recognition and the differences between LRNN and CHMM classification in further tests. For the tests, feature vectors extracted from the test material were manipulated before scoring. Each feature value which did not exceed a certain threshold value was set to zero. The recognition rates were measured as a function of the threshold. The results are shown in Fig. 2. It can be seen that the distinct peaks in the representation of the speech signals are the most relevant information for LRNN. A recognition rate above 90% is maintained even if 80% of the feature values are set to zero. CHMM recognition, on the other hand, needs all information encoded in the features including the low values between distinct peaks which are more distorted in background noise,

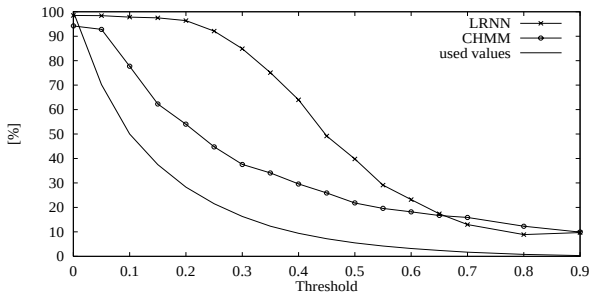

Figure 2. Recognition Rates for LRNN and CHMM as function of threshold for the values of PEMO features. All feature values below the threshold were set to zero.

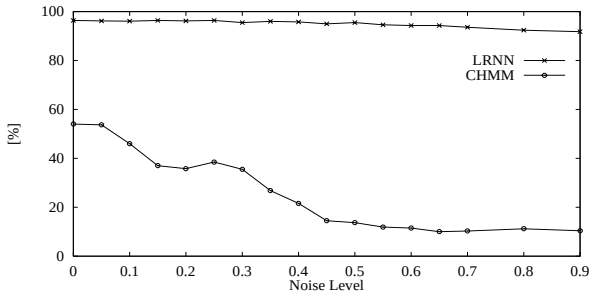

Figure 3. Recognition Rates for LRNN and CHMM as function of noise level between the distinct peaks of PEMO representation. See text.

as can be seen in Fig. 1. In a second test, each feature value which did not exceed a threshold value of 0.2 (70% of all feature values) was set to a random value between 0 and x. The recognition rates were measured as a function of x. The results are shown in Fig. 3. Even if the feature values below the threshold are heavily disturbed, LRNN is almost not affected in its performance. Due to recurrent connections LRNN is able to classify information which is distributed over time. A pattern can be recognized even if the space "in between" the prominent peaks is heavily disturbed. CHMM-based recognition, on the other hand, scores each single time frame without regarding temporal context. Distortions between the relevant information of the sparse and distinct PEMO coding then have a strong impact on the recognition performance.

5. EXPERIMENT II

The envelope modulation spectrum of speech typically shows a broad peak between 3-8 Hz modulation frequency which origins from the average rate of phonemes and articulator movement [5]. In human speech perception, analysis of low modulation frequencies appears to play a major role. In a recent study on the intelligibility of temporally-smeared speech it was found that modulations at rates above 16 Hz are not required for speech intelligibility [6].

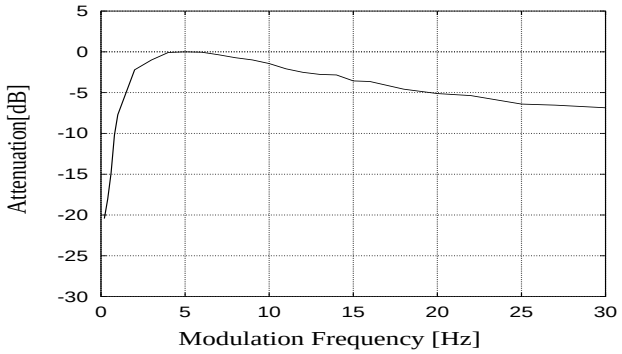

Figure 4. Envelope modulation transfer function of PEMO in the low modulation frequency range. Very slow envelope fluctuations are attenuated by the steady-state compression, fast fluctuations are suppressed by the 8 Hz low pass filter at the end of PEMO processing

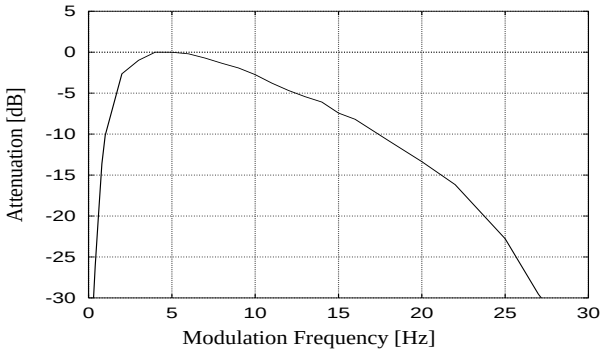

Figure 5. Envelope modulation transfer function of PEMO with enhanced modulation band pass filtering (FILT)

PEMO processing has an inherent envelope modulation band pass filter which attenuates very slow changes as well as quick changes in the envelope due to the adaptive compression and the 8 Hz low pass filter, respectively. PEMOs inherent modulation transfer function is shown in Fig. 4. The maximum envelope modulation transmission of the model can be found at modulation frequencies around 6 Hz. Thus, the influence of very slow or fast fluctuating noise is weakened. Hermansky *et al.* introduced modulation bandpass filtering of spectral components for speech recognition in noise and speech enhancement (RASTA [2]), but with much steeper filter functions than PEMO. We applied this kind of bandpass filter on PEMO feature vectors, thus increasing PEMOs inherent filtering as shown in Fig. 5 (FILT). Recognition experiments were carried out in different test conditions. The results are shown in Table 1 (last row). It can be seen that additional modulation filtering lessens the influence of noise on the feature vectors and allows a further increase of the recognition rate in combination witch LRNN as classifier. In additive noise at 10 dB SNR, a recognition rate of 95.1% was reached. Telephone digits could be recognized with 94.5%, even if the training was performed on telephone-filtered studio speech. (When trained on real telephone speech, a recognition rate of 97.4% was reached). The CHMM recognizer profits from enhanced modulation filtering, as well, but still the recognition rates are much lower than with the LRNN classifier.

6. CONCLUSION

Due to the ability of LRNN to exploit information which is distributed over time and to consider temporal context, it is predestinated to take advantage of PEMO processing of speech which supplies a sparse and distinct representation of the input signal. The prominent peaks of this representation are well maintained in noise and allow high recognition rates even under poor conditions. Modulation frequencies outside the range of the average envelope modulation spectrum of speech do not have to be encoded in the signal representation. Their attenuation further decreases the influence of both additive and convolutive noise and is a further step towards robust speech recognition. The computational effort for PEMO and LRNN does not rule out applications in "real" ASR systems. Current work focuses on implementing both techniques in hardware.

REFERENCES

[1] Kasper, K., Reininger, R., and Wolf, D., "Exploiting the Potential of Auditory Preprocessing for Robust Speech Recognition by Locally Recurrent Neural Networks," Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2, pp. 1223-1227, 1997

[2] Hermansky, H., and Morgan, N., "RASTA Processing of speech," IEEE Trans. Speech Audio Processing, 2, pp. 578-589, 1994

[3] Dau, T., Püschel, D., and Kohlrausch, A., "A quantitative model of the "effective" signal processing in the auditory system: II. Simulations and measurements," J. Acoust. Soc. Am., 99, pp. 3633-3631, 1996

[4] Kasper, K., Reininger, R., Wolf, D., and Wüst, H., "A Speech Recognizer Based on Locally Recurrent Neural Networks," Proc. Int. Conf. on Artificial Neural Networks, 2, pp. 15-20, 1995

[5] Houtgast, T., and Steeneken, H.J.M., "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Am., 77, pp. 1069-1077, 1985

[6] Drullman, R., Festen, J.M., and Plomp, R., "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am., 95, pp. 1053-1064, 1994