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

J. Tchorz*, K. Kasper†, H. Reiminger‡, 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.

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 and takes about 1.4 times real time on an SGI RS/5000 workstation.

LRNN recognition are biologically motivated and have been introduced in 1994 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.

4. EXPERIMENT

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 \( \text{LOG} \). The second variation of PEMO went in 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 \( \text{MOD} \). 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 \( \text{CHMM} \). 5 Gaussian mixtures per state, diagonal covariance matrices and 6 emitting states per word model were used for the experiments.
Table 1. Speaker-independent recognition rates in percent from experiments I (first three rows) and II (last row). CLN: clean speech, S/1/0: speech simulating noise added at 10 dB SNR to the test material, TEL: real telephone speech for testing.

<table>
<thead>
<tr>
<th>Method</th>
<th>CLN</th>
<th>S/1/0</th>
<th>TEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRNN</td>
<td>40%</td>
<td>20%</td>
<td>10%</td>
</tr>
<tr>
<td>CHMM</td>
<td>60%</td>
<td>40%</td>
<td>30%</td>
</tr>
</tbody>
</table>

Spoken-independent recognition rates were measured on three sets of test data: undistorted speech, CLN, speech which was distorted by additive noise, simulating noise at 20 dB SNR before frame extraction (S/1/0), and speech recorded via telephone line (TEL). The recognition rates are shown in Table 1. For LRNN, no significant difference can be observed between PEMO and MOD. With PEMO, high recognition rates are yielded in clean speech, but the rate drops to chance when the test material is disturbed by additive noise or transmission noise (TEL). For the CHMM recognizer, the three types of features allow comparable results in clean speech. In distorted speech, PEMO or MOD feature extraction helps to increase recognition rates significantly compared to fixed compression in LOG, but the rate drops to chance if the test material is read 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 even in combination with LRNN. If the prominent peaks which encode the temporal evolution of the signal are missing (MOD), non-distinct cues for LRNN recognition are left in disturbed speech.

4.2. Manipulating the threshold

We analyzed the contribution of sparse and distinct peaks in PEMO processing to robust recognition and the differences between LRNN and CHMM classification in a second test. For this test, frames 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 obtained in background noise.

5. EXPERIMENT II

The evoking modulation spectrum of speech typically shows a broad peak between 30 and 60 Hz which reflects high-pass filtering by the vocal tract and articulator movement [12]. In human speech perception, analysis of low modulation frequency appears to play a major role. In contrast, studies on the intelligibility of compressed speech found only a small effect of modulation at 30 Hz on speech intelligibility [1].

The green line shows the actual boundary between distinct peaks and the red line shows the frequency which originates from the average mix of phonemes and articulator movement [1]. In human speech perception, analysis of low modulation frequency appears to play a major role. In contrast, studies on the intelligibility of compressed speech found only a small effect of modulation at 30 Hz on speech intelligibility [1].

Figure 2. Recognition Rates for LRNN and CHMM as a 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 a function of threshold for the values of PEMO features. All feature values below the threshold were set to zero.
Due to the ability of LEHNN to exploit information which is distributed over time and to consider temporal context, it is recommended to take advantage of LEHNN 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 responses outside the range of the average envelope modulation spectrum of speech do not have to be accounted in the signal representation. Their attenuation further decreases the influence of both additive and convolutive noise and is a further step towards robust signal recognition. The computational effort for LEHNN and LEHNN does not rule out applications in real-time ASR systems. Current work focuses on implementing both techniques in hardware.

### REFERENCES


### Table 1

<table>
<thead>
<tr>
<th>Modulation Frequency [Hz]</th>
<th>Attenuation [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>-5</td>
</tr>
<tr>
<td>10</td>
<td>-10</td>
</tr>
<tr>
<td>15</td>
<td>-15</td>
</tr>
<tr>
<td>20</td>
<td>-20</td>
</tr>
<tr>
<td>25</td>
<td>-25</td>
</tr>
<tr>
<td>30</td>
<td>-30</td>
</tr>
</tbody>
</table>