EXPLOITING THE POTENTIAL OF AUDITORY PREPROCESSING FOR ROBUST SPEECH RECOGNITION BY LOCALLY RECURRENT NEURAL NETWORKS

Klaus Kasper, Herbert Reininger, and Dietrich Wolf

Institut für Angewandte Physik
Johann Wolfgang Goethe-Universität, 60054 Frankfurt, FRG
kasper,herbrein@apx00.physik.uni-frankfurt.de

ABSTRACT

In this paper we present a robust speaker independent speech recognition system consisting of a feature extraction based on a model of the auditory periphery, and a Locally Recurrent Neural Network for scoring of the derived feature vectors. A number of recognition experiments were carried out to investigate the robustness of this combination against different types of noise in the test data. The proposed method is compared with Cepstral, RASTA, and JAH-RASTA processing for feature extraction and Hidden Markov Models for scoring. The presented results show that the information in features from the auditory model can be best exploited by Locally Recurrent Neural Networks. The robustness achieved by this combination is comparable to that of JAH-RASTA in combination with HMM but without any requirement for an explicit adaptation to the noise in speech pauses.

1. INTRODUCTION

Robustness of recognition performance against additive background noise and convolutive distortions due to channel or microphone characteristics is still a serious problem in automatic speech recognition. One approach to overcome this problem is to apply a model of auditory preprocessing for extraction of noise robust feature vectors. In the context of speech recognition, the most promising models are those which take into account the findings of psychoacoustic experiments. Ideally, the features extracted with such a model represent the information contained in a speech signal in the most relevant form. Properties of human speech perception, like frequency warping, loudness sensitivity, or noise masking, can be exploited without exact knowledge about the signal processing in the auditory system.

However, the representation of speech signals delivered from a psychoacoustically oriented auditory model does not automatically lead to robust speech recognition. Additionally, an adequate technique for exploiting the information represented in the feature sequences is essentially required. Hidden-Markov-Models (HMM) and Artificial Neural Networks are recently the most promising approaches for the acoustic modeling and scoring of feature sequences. With respect to speaker independent word recognition both approaches show about the same performance with noise-free speech signals. Unfortunately, both concepts don't have an inherent robustness, i.e. are sensitive against noise in a sense that speech recognizers optimized with noise-free speech data show a dramatic decrease of their recognition performance for noisy speech

In this paper, a robust speech recognition system (SRS) for speaker independent recognition of isolated words is presented. It consists of a feature extraction based on a model of the auditory periphery, and a Locally Recurrent Neural Network (LRNN) for scoring of the derived feature vectors. A number of recognition experiments were carried out to investigate the robustness of this combination against different types of noise in the test data. The results were compared with other approaches for feature extraction and scoring techniques used in SRS.

2. AUDITORY PERCEPTION MODEL

The auditory perception model (PEMO) simulates the processing of the auditory system [1]. The speech signal is preemphasized by first-order differentiation and filtered by a gammatone filterbank with 19 bandpass filters with center frequencies from 330-4000 Hz. The bandwidths and center frequencies of the gammatone filterbank simulate the critical filters in the auditory system. Each frequency channel is half-wave rectified and lowpass filtered at 1 kHz for envelope extraction. This simulates the limiting phase-locking for auditory nerve fibers at high frequencies. The next stage of the perceptive preprocessing is a set of five consecutive adaptation loops. Each of these loops consists of a divider and a RC-low pass. The input signal is divided by the low pass filtered output signal. Thus, fast changes of the envelope are transformed linearly, whereas slow variations of the envelope are compressed by a square root law. Combining five consecutive adaptation loops approximates the logarithm of the average input fairly well. Due to this adaptive compression unit, changes in the input signal are contrasted, constant parts of the input are suppressed. The last step of the perceptive preprocessing is an 8 Hz low-pass filter at the output of the adaptation loops for each center frequency. The low-pass filter characterizes the auditory system's sluggishness in following rapid envelope fluctuations. The output of the auditory model was blocked into frames of 10 ms with 19 coefficients. In Figure 1 the waveform of an utterance of the german word *Sieben* and its representation by PEMO processing is shown. It can be seen that PEMO processing emphasizes fast changes in the waveform caused by transitions of phones. In the presence of noise this emphasized representation of speech will be

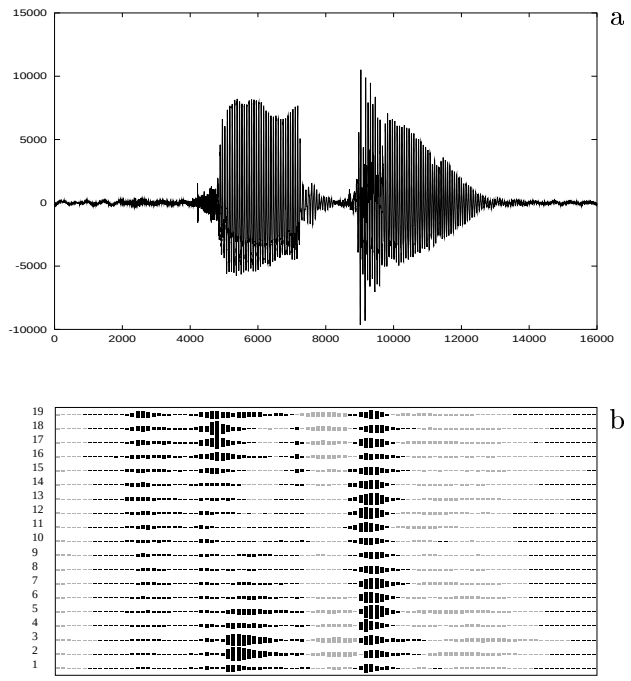

Figure 1. Waveform of an utterance of *Sieben* and its representation by PEMO processing

preserved although it may be slightly degraded.

3. LOCALLY RECURRENT NEURAL NETWORKS

Locally Recurrent Neural Networks (LRNN) are biologically motivated and have been introduced [2, 3] in order to reduce the computational complexity of fully connected recurrent neural networks. It has been shown that SRS based on LRNN achieve recognition results for isolated words and connected digits which are comparable to sophisticated HMM based systems. A LRNN consists of an input layer, a hidden layer, and an output layer. The interactions between the input and the hidden layer as well as between the hidden and the output layer are unidirectional and sparse. The recurrent connections of the hidden neurons are ending at the edges of the grid. In Figure 2 the weight values of a LRNN are shown in a Hinton Diagram. The LRNN was trained to recognize 10 digits on the basis of PEMO processing. The recognition performance of this particular LRNN and other similar designed networks trained on the basis of different kinds of features will be reported later on. The input layer consists of 95 neurons, the hidden layer of 169 neurons, and the output layer of 10 neurons. The absolute values of the weights are coded as gray scaled squares. It can be seen that each neuron of the hidden layer has connections to only one fifth of the input neurons. Moreover, the recurrent connections in the hidden layer are restricted to the four nearest neighbours of each hidden neuron. Also obvious is that connections from input to output neurons and from output back to hidden neurons do not exist. The described topology of the LRNN leads to a total amount of about 15000 weights in contrast to 50000 weights of a

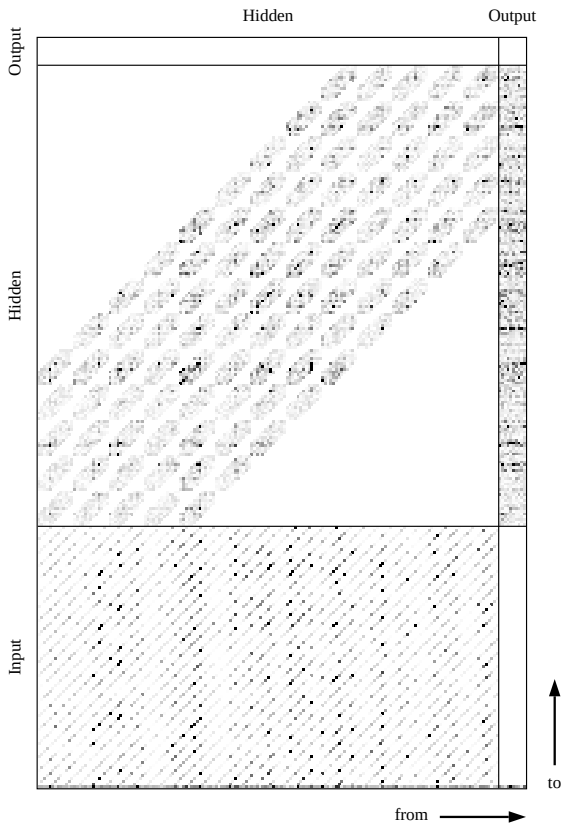

Figure 2. Weight values of a LRNN based on PEMO processing for recognition of 10 digits

fully connected recurrent network with the same amount of neurons. In simulation experiments it has been revealed that the amount of neurons can be reduced in the case of fully connected recurrent networks so that a total amount of about 30000 weights resulted. Nevertheless, the application of LRNN for recognition of isolated word leads to a tremendous reduction in computational load and complexity. Thus, a hardware realization of LRNN for speech recognition as single chip solution could be done.

LRNN are trained by truncated back-propagation through time (BPTT). 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. This furthermore, supports a single chip realization of the complete speech recognizer.

4. RECOGNITION EXPERIMENTS

4.1. Speech Data

In order to evaluate the proposed SRS consisting of perceptually-based preprocessing and LRNN we configured systems using different preprocessing modules as well as scoring techniques for speaker-independent recognition of isolated German words.

The vocabulary of the data base used in the simulation

experiments consists of the 10 German digits. For computing the parameters of the various SRS feature vectors from 100 utterances of each word spoken from different speakers were used. Speaker-independent recognition rates were measured on a set containing 100 utterances of each word from speakers not included in the training set. The robustness of speech recognition was tested with additive background noise and convolutive noise caused by changes in the microphone and the telephone channel. Three different types of noise were used: 1) white gaussian noise (WGN), 2) speech-simulating noise (SSN), which was generated from a random superposition of words spoken by a male speaker, 3) background noise recorded on a construction site (CON). The first two noise types (WGN and SSN) are stationary, i.e. their spectral shape and power do not change over time. The third noise (CON) exhibits fluctuations in both spectral shape and power. Each of these three background noises was added to the test material before feature extraction at two signal-to-noise ratios (SNR). The training was always done with clean speech.

To introduce another more realistic noisy environment, a second set of test data (TUBTEL) was applied. These data consist of 117 utterances of each digit, spoken by different persons over dialed-up public telephone lines in the Berlin area [4]. Therefore, microphone and channel characteristics strongly influenced the recorded speech signals. These telephone speech data were not used for training, but only for testing.

4.2. Feature Extraction

Four preprocessing techniques were applied to compare the robustness of the feature vectors: 1) Cepstral Coefficients (LPC-CEP) – 12 cepstral coefficients, derived from LPC parameters, and short term energy calculated every 10 ms using 32 ms Hamming windows. The temporal derivatives of these values were used as additional dynamic features. 2) RASTA processing – the speech was processed using a 20 ms Hamming window, the critical bands were transformed with a logarithmic mapping before bandpass filtering and parameterized into 9 PLP-cepstral values [5]. 3) JAH-RASTA processing – this type of processing was introduced in order to improve the robustness of RASTA processing against additive and convolutive type of noise by adapting the feature extraction to the noise level of an utterance [6]. The J value which determines the noise level adaptation is calculated from the first 125 ms of each utterance, which has to be speech-free for optimized performance. To transform the spectrum obtained from a J value corresponding to noisy speech to a spectrum processed with a J value for clean speech, mapping coefficients were calculated from J-RASTA filtered critical band outputs from a subset of the training utterances. 4) PEMO processing – the speech data was preprocessed with the auditory perception model as described above. The gammatone filterbank consisted of 19 bandpass filters with center frequencies from 330-4000 Hz. The output of the auditory model was sampled every 10 ms.

4.3. Scoring

We evaluated the proposed SRS based on LRNN by comparison with SRS based on Continuous (CHMM) or Discrete Hidden Markov Models (DHMM). In the case of the CHMM based SRS every word model consists of 8 emitting states each containing 5 gaussian mixtures with diagonal covariance matrices. This configuration leads to about 33000 parameters which have to be adjusted in the process of training. In the case of DHMM based SRS the word models consist of 8 emitting states. The feature vectors are divided into two streams, one consists of the basic feature coefficients and the other one of the derivated coefficients. For each stream a codebook with 128 vectors was trained. In the training mode about 20000 parameters had to be adjusted. But one has to take into account that the two codebooks consist of about 5000 additional parameters which had to be adjusted separately. For scoring the feature vectors during recognition mode the three nearest codebook vectors are used in a fuzzy manner.

Preliminary experiments have shown that HMM based SRS benefits from the usage of dynamic information. Therefore, so-called Delta coefficients were used together with the basis coefficients in the case of all feature extraction modules for SRS based on HMM. Moreover, it was revealed in these experiments that DHMM based SRS show a significantly higher robustness against additive and convolutive noise as CHMM based systems. These results were also found in previous experiments [7]. In the reported experiments we used therefore DHMM based SRS which are configured to use basis coefficients along with Delta coefficients in two different streams. In both HMM based SRS a word hypothesis is generated on the basis of the optimum path for every word, calculated by a standard Viterbi algorithm.

The LRNN consists of 169 neurons in the hidden layer with recurrent connections between the 5 nearest neighbours and 10 neurons in the output layer for representing the words of the vocabulary. Because of the sparse connectivity each neuron of the hidden layer is connected to only one fifth of the input neurons. As input patterns 5 consecutive non-overlapping feature vectors are used i.e. without providing explicit delta information. By gluing five feature vectors together the time span a LRNN can memorize is extended to the duration of a complete word. The number of neurons in the input layer varies between 45 and 130 because the size of the feature vectors depends on the type of preprocessing. This architecture results in about 15000 weights of the LRNN which had to be adapted in the training. By accumulating the activities of the output neurons for all patterns of an utterance the word hypothesis is calculated.

4.4. Results

In Table 1 the recognition performance for DHMM and LRNN based SRS using the different procedures for feature extraction are shown. It can be seen that the cepstral representation of speech is highly sensitive against noise, even if delta information is used as in the case of DHMM. Using RASTA processing leads to a significantly improved performance in noisy environments. In the case of a SNR of 20 dB SRS using RASTA processing show a fairly well robustness for stationary and convolutive noise. The sensitivity against changing noise characteristics and additive noise is due to the RASTA algorithm. JAH-RASTA processing overcomes these limitations by an explicit adaptation to the

Table 1. Speaker independent, isolated digit recognition rates in per cent from DHMM and LRNN based recognition systems for different types of noise

	LPC-CEP		RASTA		JAH-RASTA		PEMO	
NOISE	DHMM	LRNN	DHMM	LRNN	DHMM	LRNN	DHMM	LRNN
clean	97.9	96.6	99.2	98.9	99.0	98.2	96.7	98.1
WGN 10 dB	12.9	10.0	71.7	49.5	84.7	86.6	59.7	80.1
WGN 20 dB	52.4	11.3	94.2	86.8	96.3	94.3	87.5	95.1
SSN 10 dB	67.7	41.5	77.1	79.6	84.4	78.6	77.7	84.6
SSN 20 dB	94.0	80.1	96.1	96.0	97.0	94.0	94.5	96.5
CON 10 dB	36.0	30.0	66.7	50.5	82.5	81.4	76.0	89.4
CON 20 dB	84.2	72.9	92.5	87.5	97.3	94.3	94.3	97.2
TUBTEL	81.1	83.0	90.5	88.5	91.3	90.8	86.3	94.5

noise characteristics and level. Our experiments show that this adaptation leads to a significantly improved robustness against additive construction noise and telephone speech which are the more realistic environments. For all three so far discussed types of feature extraction DHMM and LRNN based SRS exhibit about the same recognition performance and sensitivity against the various noise types. In combination with PEMO processing, DHMM achieve a comparable robustness against WGN and SSN as DHMM in combination with RASTA processing. Against CON and TUBTEL the DHMM based SRS using PEMO show a slightly better robustness than the combination of DHMM with RASTA processing. But in all cases the combination with JAH-RASTA delivered the highest robustness among the DHMM based SRS.

In contrast, LRNN based SRS reach for all types of noise the highest performance in combination with PEMO. Obviously, scoring with LRNN takes the most advantage out of the perceptive representation of speech gained from the auditory model. In comparison to the best DHMM based SRS, which is using JAH-RASTA processing, the here presented SRS consisting of PEMO processing and LRNN show in the case of WGN and SSN about the same robustness, but in the case of CON and TUBTEL a significantly higher robustness. It is important to notice that PEMO works better for non-stationary noise like CON and in realistic environments like TUBTEL, although no explicit adaptation to the noise environment is done. This result demonstrates the outstanding capability of PEMO in combination with LRNN to exploit psychoacoustical findings for robust speech recognition. Furthermore, it is remarkable that the LRNN based SRS is able to exploit the inherent robustness of PEMO with a significantly smaller number of parameters than even DHMM based SRS.

5. CONCLUSIONS

The combination of perceptive preprocessing and LRNN for scoring leads to significantly higher recognition rates than systems with cepstral or RASTA coefficients as feature vectors. Adaptive JAH-RASTA processing in combination with DHMM gives comparable results for stationary and convolutive noise, but some restrictions have to be accepted: the beginning of each test utterance is assumed to be speech-free to calculate an estimate of the noise power. If the power of the background noise changes during the utterance, the calculated J-value is no longer valid. Perceptive preprocessing on the other hand is independent on assumptions about the noise level, no parts of the input signal are assumed to be speech–free, no mapping between feature vectors from quiet and noisy utterances is needed. If the noise level changes during the utterance, the preprocessing adapts instantaneously. Obviously, LRNN are able to model and exploit the characteristics of this type of features. Thus, the proposed combination of PEMO with LRNN achieves a highly robust recognition performance. Current investigations are concentrated on a detailed analysis of the interactions between the auditory preprocessing and the neural modeling of the features.

Acknowledgement

We would like to thank Birger Kollmeier and Jürgen Tchorz from Oldenburg University who introduced and realized PEMO processing and Peter Ratuschni from Frankfurt University who carried out the experiments on DHMM.

REFERENCES

[1] Dau, T., Püschel, D., and Kohlrausch, A. *A quantitative model of the 'effective' signal processing in the auditory system: I. Model Structure. J. Acout. Soc. Am.*, 1996. in press.

[2] Kasper, K., Reininger, H., Wolf, D., and Wüst, H. *A Speech Recognizer Based on Locally Recurrent Neural Networks.* In *Proc. Int. Conf. on Artificial Neural Networks*, vol. 2, pp. 15–20, Paris, 1995.

[3] Kasper, K., Reininger, H., and Wüst, H. *Strategies for Reducing the Complexity of a RNN Based Speech Recognizer.* In *Proc. IEEE Conf. on Acoust., Speech, and Signal Processing*, vol. 6, pp. 3354–3357, Atlanta, 1996.

[4] Schürer, T., Fellbaum, K., Ahrling, S., Hardt, D., Klaus, H., Mengel, A., Sahm, O., and Suhardi, I. *TUBTEL - Eine deutsche Telefon-Sprachdatenbank.* In *Studientexte zur Sprachkommunikation, Heft 12*, pp. 183–187, 1995.

[5] Hermansky, H. and Morgan, N. *Rasta Processing of Speech. IEEE Transactions on Speech and Audio Processing*, 2:578–589, 1994.

[6] Koehler, J., Morgan, N., Hermansky, H., Hirsch, G., and Tong, G. *Integrating RASTA-PLP into Speech Recognition.* In *Proc. IEEE Conf. on Acoust., Speech, and Signal Processing*, vol. I, pp. 421–424, Adelaide, 1994.

[7] Reininger, H. Stochastische und neuronale Konzepte zur automatischen Spracherkennung. Hector, Frankfurt am Main, 1994.