ELSEVIER

Speech Communication 34 (2001) 75–91

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Combining speech enhancement and auditory feature extraction for robust speech recognition

Michael Kleinschmidt *, Jürgen Tchorz, Birger Kollmeier

AG Medizinische Physik, Universität Oldenburg, 26111 Oldenburg, Germany

Abstract

A major deficiency in state-of-the-art automatic speech recognition (ASR) systems is the lack of robustness in additive and convolutional noise. The model of auditory perception (PEMO), developed by Dau et al. (T. Dau, D. Püschel, A. Kohlrausch, J. Acoust. Soc. Am. 99 (6) (1996) 3615–3622) for psychoacoustical purposes, partly overcomes these difficulties when used as a front end for automatic speech recognition. To further improve the performance of this auditory-based recognition system in background noise, different speech enhancement methods were examined, which have been evaluated in earlier studies as components of digital hearing aids. Monaural noise reduction, as proposed by Ephraim and Malah (Y. Ephraim, D. Malah, IEEE Trans. Acoust. Speech Signal Process. ASSP-32 (6) (1984) 1109–1121) was compared to a binaural filter and dereverberation algorithm after Wittkop et al. (T. Wittkop, S. Albani, V. Hohmann, J. Peissig, W. Woods, B. Kollmeier, Acustica United with Acta Acustica 83 (4) (1997) 684–699). Both noise reduction algorithms yield improvements in recognition performance equivalent to up to 10 dB SNR in non-reverberant conditions for all types of noise, while the performance in clean speech is not significantly affected. Even in real-world reverberant conditions the speech enhancement schemes lead to improvements in recognition performance comparable to an SNR gain of up to 5 dB. This effect exceeds the expectations as earlier studies found no increase in speech intelligibility for hearing-impaired human subjects. © 2001 Elsevier Science B.V. All rights reserved.

Zusammenfassung

Die mangelnde Robustheit moderner Systeme zur automatischen Spracherkennung gegenüber additiven und konvolutiven Störungen ist eines der drängensten Probleme aktueller Forschung. Das Perzeptionsmodell nach Dau et al. (T. Dau, D. Püschel, A. Kohlrausch, J. Acoust. Soc. Am. 99 (6) (1996) 3615–3622), welches ursprünglich für psychoakustische Anwendungen konzipiert wurde, kann als auditorische Vorverarbeitung zu einer robusteren Erkennungsleistung beitragen. Um die Klassifikationsleistung dieses gehörbasierten Erkennungssystems weiter zu erhöhen, wurden verschiedene Methoden zu Störgeräuschunterdrückung untersucht, welche in der Vergangenheit als Komponenten digitaler Hörgeräte evaluiert wurden. Verglichen wurde das monaurale Verfahren zur Störgeräuschreduktion nach Ephraim and Malah (Y. Ephraim, D. Malah, IEEE Trans. Acoust. Speech Signal Process. ASSP-32 (6) (1984) 1109–1121) mit dem binauralen Filter und Enthallungsalgorithmus nach Wittkop et al. (T. Wittkop, S. Albani, V. Hohmann, J. Peissig, W. Woods, B. Kollmeier, Acustica United with Acta Acustica 83 (4) (1997) 684–699). In reflexionsarmer Umgebung bewirkten beide Algorithmen eine Erhöhung der Erkennungsleistung, entsprechend einer Verbesserung des Signal-Rausch-Abstands um bis zu 10 dB für alle untersuchten Störgeräusche, während die Ergebnisse in Ruhe nicht beeinträchtigt wurden. Selbst in realer, verhallter Umgebung erreichten die Störunterdrückungsverfahren

[*] Corresponding author. Tel.: +49-441-798-3146; fax: +49-441-798-3902.
E-mail address: michael@medi.physik.uni-oldenburg.de (M. Kleinschmidt).

PII: S0167-6393(00)00047-9

Verbesserungen der Erkennungsleistung vergleichbar einem um bis zu 5 dB günstigeren SNR. Diese Ergebnisse übertreffen die Erwartungen, da in früheren Untersuchungen für schwerhörige Versuchspersonen mit digitalen Hörgeräten keine Erhöhung der Sprachverständlichkeit gefunden werden konnte.

Résumé

Un des problèmes les plus urgents de la recherche actuelle des systèmes de reconnaissance de la parole automatique est leur robustesse déficiente envers du bruit additif et la réverbération. Le modèle de perception auditive (PEMO) réalisé par Dau et al. (T. Dau, D. Püschel, A. Kohlrausch, J. Acoust. Soc. Am. 99 (6) (1996) 3615–3622) pour une application dans le domaine psychoacoustique peut partiellement surmonter ces difficultés, s'il est appliqué comme prétraitement pour la reconnaissance de la parole automatique. Afin de perfectionner la performance de ce système auditif de reconnaissance de la parole automatique en bruit d'environnement, plusieurs méthodes de débruitage de la parole furent examinées, qui étaient évaluées comme composants des prothéses auditives dans le passé. La réduction monaurale de bruit comme proposée par Ephraim and Malah (Y. Ephraim, D. Malah, IEEE Trans. Acoust. Speech Signal Process. ASSP-32 (6) (1984) 1109–1121) fut comparée avec le filtre binaural et l'algorithme de réverbération d'après Wittkop et al. (T. Wittkop, S. Albani, V. Hohmann, J. Peissig, W. Woods, B. Kollmeier, Acustica United with Acta Acustica 83 (4) (1997) 684–699). Tous les deux algorithmes de réduction de bruit améliorent la performance de reconnaissance correspondant à une amélioration de jusqu'à 10 dB de rapport signal/bruit pour tous les bruits d' environnement étudiés, pendant que les résultats obtenus pour la parole présentée sans bruit ne furent pas diminués considérablement. Même dans un environnement réel sans réverbération ces méthodes de réduction de bruit améliorent la performance de reconnaissance correspondant à une amélioration de jusqu'à 5 dB de rapport signal/bruit. Ces résultats dépassent les prévisions, parce que dans des anciennes études on n'avait pas obtenu une augmentation de l'intelligibilité de la parole pour des patients avec des déficiences auditives.

Keywords: Robust speech recognition; Perceptive modeling; Auditory front end; Speech enhancement

1. Introduction

A major problem of most automatic speech recognition (ASR) systems is their unsatisfactory robustness in noise. Several researchers proposed front ends which simulate different processing stages of the auditory system to overcome this problem, as human 'feature extraction' leads to very robust speech understanding in noise (Ghitza, 1988; Seneff, 1988). In recognition experiments, however, these auditory-based front ends often yield only small or no improvements compared to standard front ends, or require high computational costs (Jankowski et al., 1995). A further approach of auditory feature extraction is investigated here. It is based on a model of the auditory periphery (PEMO), which was originally developed by Dau et al. (1996a) to predict human performance in typical psychoacoustical masking experiments, but was also applied to different tasks in the field of speech processing (Hansen and Kollmeier, 1997; Holube and Kollmeier, 1996). It has been shown that using PEMO as a front end for automatic speech recognition systems results in additional robustness compared to standard mel-frequency cepstral coefficient (MFCC) front ends (Tchorz and Kollmeier, 1999), especially when applying locally recurrent neural networks (LRNN) as classifiers for isolated word recognition tasks (Tchorz et al., 1997).

Another method to overcome the lack of robustness observed for state-of-the-art ASR systems is to enhance the incoming time signal before feature extraction. A number of single-channel noise reduction algorithms have been examined as pre-processing steps for ASR (recent work e.g. Mine et al., 1996; Fischer and Stahl, 1999; Gelin and Junqua, 1999; Hermus et al., 1999; Vizinho et al., 1999). Since MFCC-based recognition systems are prone to degradation of performance on clean speech when combined with speech enhancement schemes (Kermorvant and Morris, 1999; Wilmers and Strube, 1999), the robustness of a given ASR system against distortions introduced by noise reduction is of major concern. Multi-channel approaches towards speech enhancement for ASR often consist of physically large microphone arrays (Kiyohara et al., 1997; Omologo et al., 1997; Bitzer et al., 1999). A special type of multi-channel processing is the *binaural* approach, i.e., a two channel approach, that assumes two microphones positioned in the 'ears' of a dummy head [1] or probe microphones close to the ears of a real person. Since this approach allows the simulation of the binaural signal processing and noise reduction usually present in normal-hearing listeners, it has attracted much attention in the area of auditory modeling (Durlach, 1972; Colburn, 1996; Blauert, 1997; Zerbs et al., 1999), noise reduction for hearing-impaired listeners (Kollmeier et al., 1993; Peissig and Kollmeier, 1997; Wittkop et al., 1997) and ASR (Bodden and Anderson, 1995; Francis and Anderson, 1997; Kleinschmidt et al., 1998).

This paper describes the benefit the combined PEMO/LRNN system may gain by employing several methods of speech enhancement, both monaural and binaural. Single-channel noise reduction algorithms such as the minimum mean square error (MMSE) short-term spectral amplitude (STSA) estimator (Ephraim and Malah, 1984) rely on temporal windows in which speech is absent to reestimate the quasi-stationary noise spectrum. Two channel algorithms in general require more technical effort, but have the possibility of directional filtering and dereverberation by exploiting the differences in phase and level between the two signals recorded at the left- and the right-hand side of a head-like object. Both types of noise reduction algorithms have been applied to the task of increasing speech intelligibility for hearing-impaired listeners (Marzinzik and Kollmeier, 1999; Wittkop et al., 1999) – but only showed a limited benefit. Although the signal-to-noise ratio (SNR) is improved by monaural and binaural speech enhancement in certain laboratory experiments, the algorithms are of limited use for humans in realistic acoustic environments. In most situations, speech intelligibility was not significantly improved, or even degraded. Some benefits could be observed in terms of 'ease of listening' and listening fatigue. One important reason for these limited benefits is that the algorithms have been used at comparatively unfavorable SNRs, where normal-hearing listeners still understand speech quite well, while impaired listeners have tremendous difficulties. ASR systems, on the other hand, have even more problems with additive noise than hearing-impaired listeners, since their performance declines at much more favorable SNRs, where noise reduction schemes yield a higher benefit. It is therefore worthwhile to combine the noise reduction strategies primarily developed for digital hearing aids with robust ASR systems to achieve an even better performance in noise. In this paper, monaural and binaural algorithms therefore have been tested for a number of types of noise and SNRs, keeping a constant PEMO front end and LRNN recognizer. The aim of this study was, on the one hand, to obtain a more robust speech recognition system and, on the other hand, an 'objective' evaluation of the speech enhancement schemes.

2. Auditory model

The model of auditory perception (PEMO) by Dau et al. (1996a) was designed as a model of the 'effective' signal processing that takes place in the auditory periphery transforming the acoustic signal into its 'internal representation'. It quantitatively accounts for a number of psychoacoustical experiments carried out with human subjects (Dau et al., 1996b, 1997), e.g. spectral and forward masking, temporal integration and modulation perception. In addition, this model has been successfully applied to the task of objective speech quality measurement (Hansen and Kollmeier, 1997), speech intelligibility prediction in noise (Wesselkamp, 1994) and in hearing-impaired listeners (Holube and Kollmeier, 1996; Derleth, 1999).

[1] Or alternatively to the left and right of a roughly head-sized and head-shaped object.

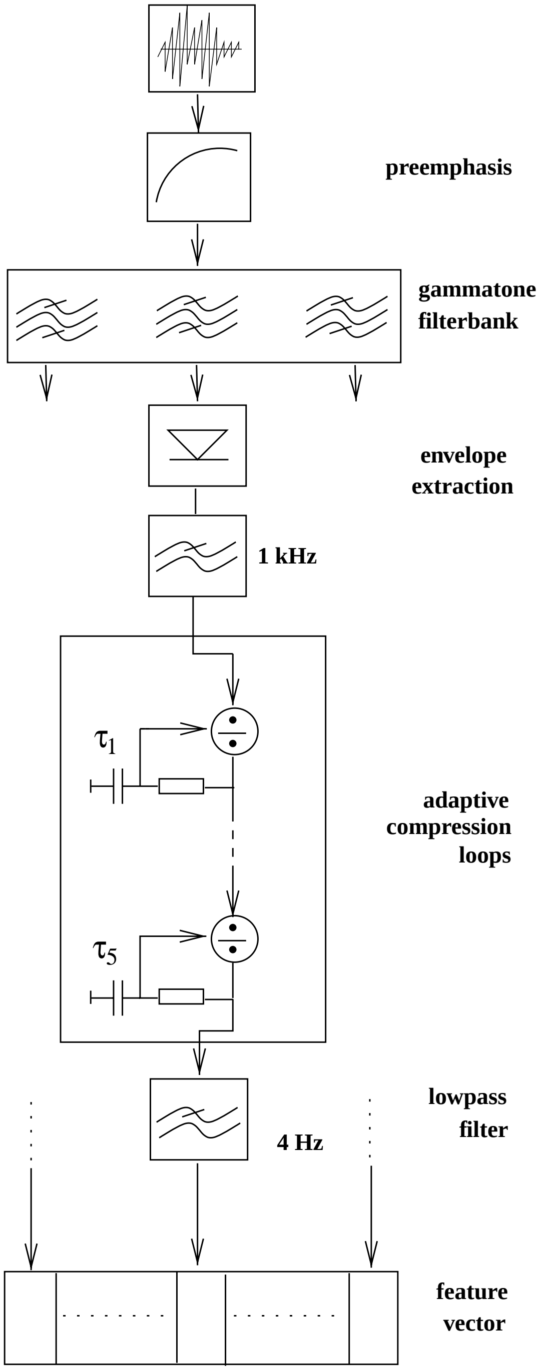

Fig. 1. Processing stages of the auditory model (PEMO).

In Fig. 1, the processing stages of the auditory model are shown. The first processing step is a pre-emphasis of the input signal with a first-order high pass filter. This flattens the typical spectral tilt of speech signals and reflects the transfer function of the outer ear. The pre-emphasized signal is then filtered by a gammatone filterbank (Patterson et al., 1987) using 19 frequency channels equally spaced on the ERB scale with center frequencies ranging from 0.3 to 4 kHz. The impulse responses of the gammatone filterbank are similar to the impulse responses of the auditory system found in physiological measurements. After gammatone filtering, each frequency channel is halfwave-rectified and first-order low pass filtered with a cut-off frequency of 1 kHz for envelope extraction, which reflects the limiting phase-locking for auditory nerve fibers above 1 kHz. Amplitude compression is performed in a subsequent processing step. In contrast to conventional bank-of-filters front ends, the amplitude compression of the auditory model is not static (e.g., instantaneously logarithmic) but adaptive, which is realized by an adaptation circuit consisting of five consecutive non-linear adaptation loops. Each of these loops consists of a divider and an RC low pass filter with an individual time constant ranging from 5 to 500 ms. Changes in the input signal like onsets and offsets are emphasized, whereas steady-state portions are compressed. Thus, the dynamical structure of the input signal is taken into account over a relatively long period of time. Short-term adaptation including enhancement of changes and temporal integration is simulated and allows a quantitative prediction of important temporal effects in auditory perception.

The last processing step of the auditory model is a first-order low pass filter with a cut-off frequency of 4 Hz. It attenuates fast envelope fluctuations of the signal in each frequency channel. Suppression of very slow envelope fluctuations by the adaptation loops and attenuation of fast fluctuations by the low pass filter results in a band pass characteristic of the amplitude modulation transfer function of the auditory model with a maximum at about 4 Hz (see Fig. 2). This corresponds well to the average modulation spectrum of speech, which also has its maximum at around 4 Hz. An extension of the model (not used here) replaces the final low pass filter by a bank of modulation band pass filters (Dau et al., 1997). The output of the auditory model is downsampled to a rate of 100 feature

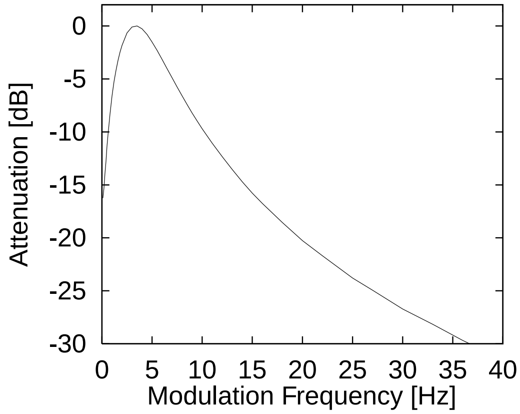

Fig. 2. Modulation transfer function of PEMO. Due to the non-linear nature of the adaptation loops the transfer function is signal-dependent. The values plotted here are calculated using an amplitude modulated sinusoidal carrier at 1 kHz.

vectors per second to serve as input to the recognizer.

3. Digit recognition experiments with PEMO front end

In this section, results from isolated word recognition experiments are introduced making use of PEMO auditory feature extraction *without* further speech enhancement. While the robustness of the PEMO front end had been documented elsewhere (Tchorz et al., 1997; Kasper et al., 1997; Tchorz and Kollmeier, 1999; Kasper and Reininger, 1999), the results of this section are intended to serve as a baseline for the following experiments with noise reduction algorithms.

3.1. Setup

A number of speaker-independent, isolated digit recognition experiments in different types of additive noise were carried out to evaluate the robustness of the auditory-based representation of speech quantitatively. The speech material for training the word models and scoring was taken from the ZIFKOM database of Deutsche Telekom AG. Each German digit was spoken once by 200 different speakers (100 males, 100 females). The speech material was equally divided into two parts for training and testing, each consisting of 1000 utterances by 50 male and 50 female speakers. Training of the word models was always performed on clean digits only. Testing was performed on clean and on distorted digits. For distortion, two types of noise were added to the utterances with SNRs between 15 and −10 dB: (a) noise which was generated from a random superposition of phonetically balanced single words from a male speaker (Sotscheck noise [2]), and (b) unmodulated speech shaped noise (CCITT G.227), with a spectrum similar to the long-term spectrum of speech.

As control front end, MFCC were examined, which are widely used in common ASR systems. The FFT-based coefficients were calculated from Hamming-windowed, pre-emphasized 32 ms segments of the input signal with a frame period of 10 ms. In our experiments, each Mel cepstrum feature vector contained 26 features (12 coefficients, log energy, and the respective first temporal derivatives).

Two different recognizers were taken for training and testing: (1) a standard continuous-density HMM recognizer with five Gaussian mixtures per state, diagonal covariance matrices and six emitting states per word model, and (2) an LRNN with three layers of neurons (95 or 130 input, 225 hidden, and 10 output neurons). Hidden layer neurons have recurrent connections to their 24 nearest neighbors. The input matrix consisted of 5 times the PEMO output vector with 19 elements, glued together in order to allow the network to memorize the whole time sequence of input matrices. In the case of using the MFCC front end the input layer consisted of five times 26 input neurons. For training, the back-propagation-through-time algorithm was applied in 200 iterations (see (Kasper et al., 1995) for a detailed description). In total, four different combinations of front ends and recognizers were compared to each other: MFCC/CHMM, MFCC/LRNN, PEMO/CHMM and PEMO/LRNN.

[2] See (Kollmeier et al., 1988).

3.2. Results

The speaker-independent digit recognition rates in clean speech and in additive noise obtained with the different combinations of front ends and recognizers are shown in Fig. 3. The results for CCITT speech shaped noise and for Sotscheck noise are shown on top and bottom, respectively. The recognition rates in percent are plotted as a function of the SNR in dB. In clean speech, all combinations yield similar recognition rates (see Table 1).

Table 1
Speaker-independent isolated word recognition rate in % on clean test data for different combinations of front end and classification tool with Ephraim–Malah speech enhancement (EM) and no processing (no)

	no	EM
PEMO–CHMM	97.6	97.4
PEMO–LRNN	98.0	97.6
MFCC–LRNN	98.0	97.7
MFCC–CHMM	98.7	97.7

In additive noise, the performance diverges. With a HMM recognizer, both front ends yield comparable results. PEMO works slightly better in Sotscheck noise, and about as good as MFCC features in CCITT noise. With the neural network as classifier, however, the choice of the front end is essential for the recognition rates. Cepstral coefficients yield only poor results in noise with the LRNN recognizer, as already reported in earlier studies by Kasper et al. (1997). When combined with the LRNN recognizer PEMO features provide a useful improvement in robustness when compared with the other combinations tested.

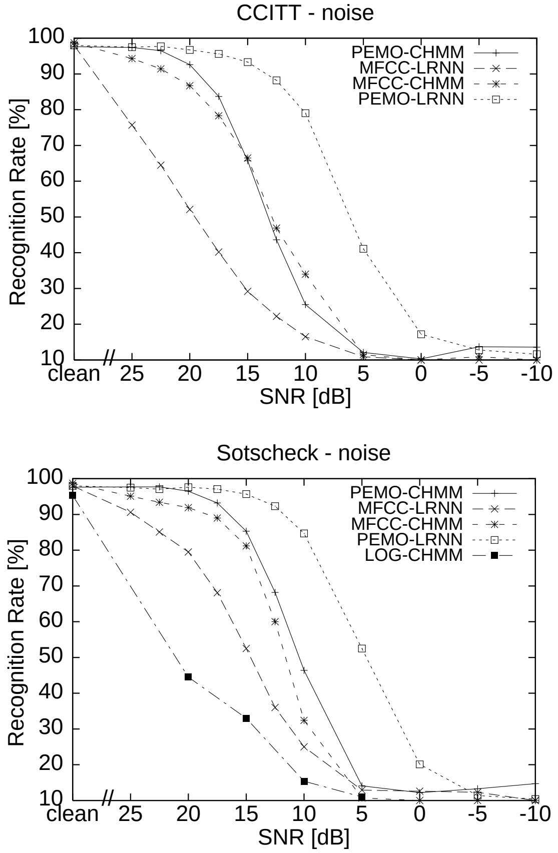

Fig. 3. Speaker-independent, isolated digit recognition rates in CCITT noise (top) and Sotscheck noise (bottom) as function of SNR for different combinations of front ends (MFCC and PEMO) and recognizers (CHMM and LRNN). The data points for condition LOG-CHMM are taken from Tchorz and Kollmeier (1999) and will be discussed in Section 7.

Tchorz et al. (1997) found that the distinct peaks in the representation of speech signals are the most relevant information for the LRNN recognizer. A recognition rate above 90% is maintained even if the 80% lowest feature values are set to zero. HMM recognition, on the other hand, shows degraded performance in that experiment. As the threshold for manipulating the features increases, the recognition rate drops rapidly. It seems as if HMM recognition exploits all information encoded in the features, including the low values between distinct peaks. These are the parts in the representation which are more distorted in background noise, as can be seen from Fig. 4, where PEMO processing is demonstrated without and with the presence of background noise.

While the LRNN seems to benefit from the sparse representation of PEMO features, this might by a problem for HMM recognizers. Also the non-diagonal elements of the covariance matrices of PEMO features are not negligibly small. As reported by Kasper and Reininger (1999) the performance of PEMO combined with CHMM recognizers can be further improved by applying a

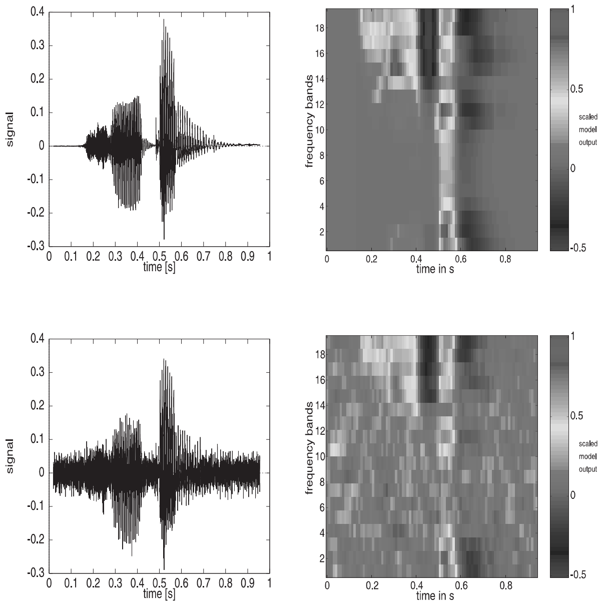

Fig. 4. Examples for PEMO processing of speech. Top: time signal of an utterance of the German digit "sieben" (left) and representation after processing (right). Bottom: same utterance disturbed with CCITT noise at 5 dB SNR and representation after processing.

cepstrum-like transformation to the features (thereby obtaining so called PEMO–CEP features). The resulting recognition scores are comparable to the performance of PEMO/LRNN.

4. Digit recognition experiments with monaural speech enhancement and PEMO front end

In order to further increase the robustness of the recognition system a single-channel speech enhancement method was added to the experimental setup.

4.1. Setup

The MMSE STSA estimator as proposed by Ephraim and Malah (1984) applies a statistically derived optimal gain to the spectral components. The gain is calculated using estimates of a-posteriori and a-priori SNR (the so-called 'decision directed approach'). This algorithm leads to an audible reduction of additive background noise without distorting the speech signal or producing 'musical tone' artifacts (Cappé, 1994). As with most single-channel noise reduction algorithms, an estimate of the noise spectrum is required. The estimate has to be updated if the additive noise is only quasi-stationary for certain time intervals. Marzinzik and Kollmeier (1999) developed a combination of the Ephraim–Malah scheme and an algorithm to automatically update the noise estimate for use in digital hearing aids. Another study focused on the usefulness of a number of variants of the Ephraim–Malah algorithm regarding its application in robust speech recognition (Kleinschmidt et al., 1999) and showed that the original filter performed better in the given setup than slightly different variants (Ephraim and Malah, 1985), that account for the uncertainty of signal presence or use logarithmic spectral amplitude values.

The experimental setup resembles the one described in Section 3 for the previous experiments and is shown in Fig. 5. The disturbed time signal is filtered by the Ephraim–Malah speech enhancement algorithm according to Marzinzik and Kollmeier (1999) before feature extraction. In addition

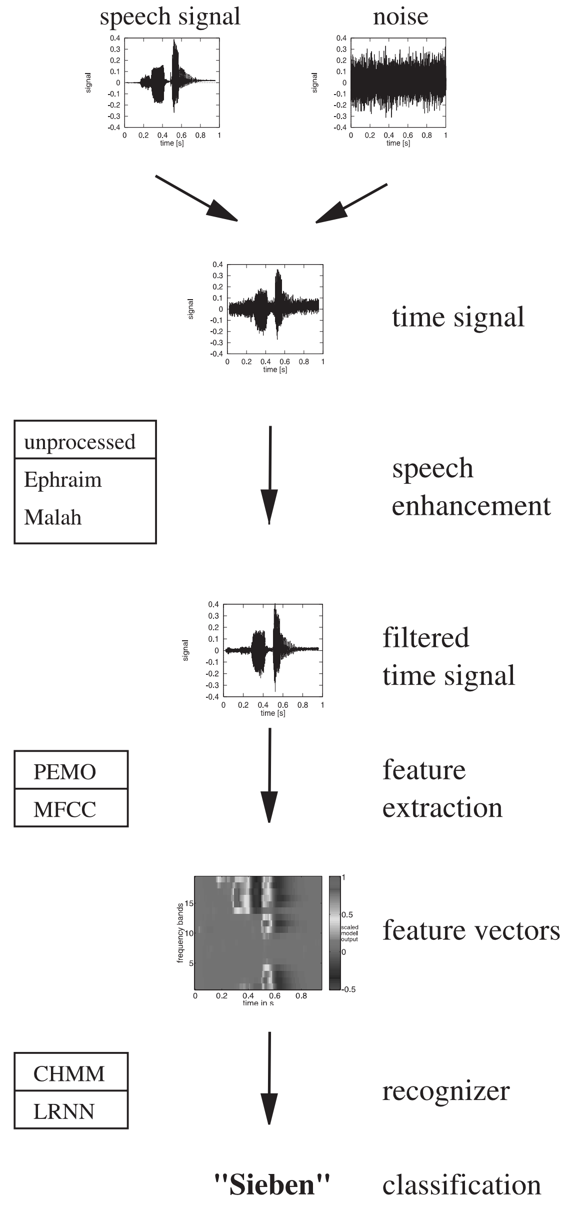

Fig. 5. Setup of isolated digit recognition experiment. Noise was added to the speech signal, which was than filtered, preprocessed and finally classified.

to Sotscheck and CCITT noise, construction site noise [3] and white Gaussian noise have been used in this comparison. The first 50 ms of each signal file were regarded as noise and therefore supplied an initial estimate of the noise spectrum. As isolated digits were used, automatic noise updating did not have to be applied necessarily, but was nevertheless included, since this is indispensable for any application in realistic environments. As above the training was carried out on one half of the dataset. The training data was left clean and unprocessed.

4.2. Results

The speaker-independent digit recognition rates in clean speech and in additive CCITT noise obtained with the different combinations of front ends and recognizers using monaural speech enhancement are shown in Fig. 6. Again, the recognition rates in percent are plotted as a function of the SNR in dB.

The performance on clean test data shows no significant degradation for all combinations of front ends and recognizers, except for MFCC/CHMM, where the recognition rate drops by one percent in total (see Table 1). This result is another hint for a rather 'gentle' noise reduction by the Ephraim–Malah algorithm. Earlier studies (Kleinschmidt et al., 1999; Wilmers and Strube, 1999) and yet unpublished experiments have shown the

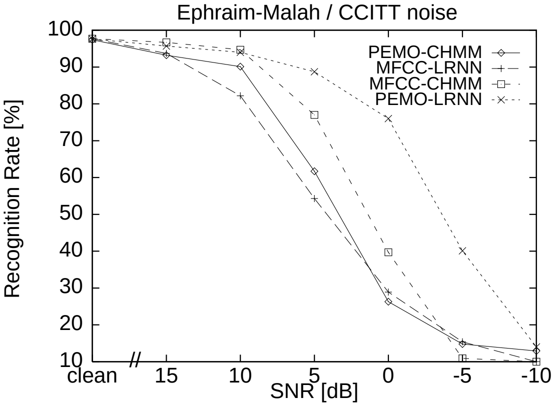

Fig. 6. Speaker-independent, isolated digit recognition rates in CCITT noise as function of SNR applying Ephraim–Malah speech enhancement for different combinations of front ends (MFCC and PEMO) and recognizers (CHMM and LRNN).

[3] Siemens Audiologische Technik 1992: CDROM for fitting and testing of hearing programs.

adverse effect of other noise reduction schemes on clean speech error rates, especially for MFCC-based systems. As an overall result the robustness of PEMO-based recognition systems was found to be higher than for MFCC front ends, not only against additive noise, but also the artifacts of speech enhancement processing.

By comparing Fig. 6 with Fig. 3 it becomes obvious that all combinations of front ends and recognizers show improved robustness in additive CCITT noise when applying the Ephraim–Malah monaural speech enhancement to the disturbed time signal before feature extraction. PEMO/LRNN is still the most robust with an effective gain of 8 dB (at 90% level) compared to no speech enhancement (c.f. Fig 3), or a gain of 50% in total recognition rate at 5 dB SNR level. The following experiments are restricted to the PEMO/LRNN recognition system as this combination appears to be the most promising one.

The speaker-independent digit recognition rates of the PEMO/LRNN combination in different types of additive noise are given in Fig. 7. The results for unprocessed and Ephraim–Malah filtered signals are located on top and on bottom, respectively.

The robustness against noise of the PEMO/LRNN recognition system significantly depends on the type of background noise added. At most SNR levels white noise seems to have the least effect on recognition performance compared to construction site noise or Sotscheck noise. Adding speech shaped CCITT noise results in the lowest recognition rates. Both CCITT and Sotscheck noise have a smooth spectrum which is similar to the long-term spectrum of speech. In contrast to CCITT, Sotscheck and in particular construction site noise are more modulated types of noise, the latter with high spectral energies at very low and very high frequencies. The results indicate that spectral distribution is a more important factor for the disturbance of speech recognition performance than modulation, at least when moderate modulation depths are compared.

It can be clearly seen in Fig. 7 that major improvements in robustness are obtained not only for CCITT noise but also for all other types of noise. This effect is less true for construction site noise, with its rather non-stationary nature. This effect is not unexpected, as the noise reduction scheme assumes temporary stationary noise while speech is active.

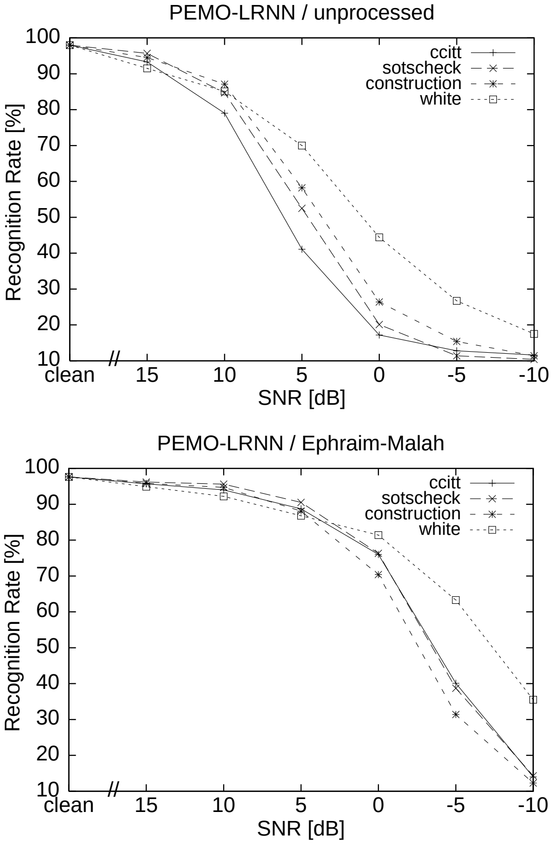

Fig. 7. Speaker-independent, isolated digit recognition rates as function of SNR with no speech enhancement (top) and Ephraim–Malah speech enhancement (bottom) for different types of noise.

5. Digit recognition experiments with binaural speech enhancement and PEMO front end

While single-channel noise suppression is based on the assumption that the noise signal is stationary, multi-channel noise reduction methods, in theory, allow for a separation of different sound sources based on spatial direction alone. In addition, unwanted reverberation can be suppressed.

This is true especially when taking advantage of the directional characteristics of a human (or dummy) head. In this section, a directional filter algorithm is examined as a pre-processing step for the digit recognition system described above.

5.1. Setup

A two-channel algorithm for the use in binaural digital hearing aids has been proposed by Kollmeier et al. (1993), Peissig (1993) and Wittkop et al. (1997). Differences in amplitude and phase between left and right input channel frequency components are used for a directional filter. Also, the interaural coherence function serves as a basis for dereverberation. A third component has been added later (Wittkop et al., 1999) for suppression of single jammer sources. The effect on speech intelligibility has been evaluated using audiometric sentence tests (Wittkop et al., 1999). While the suppression of noise was clearly audible in informal listening tests, no significant increase of speech intelligibility could be found averaged over a number of hearing-impaired subjects. As for the monaural algorithm, however, the subjective personal preferences tended towards the filtered signals.

In a previous study, virtual acoustics was used to examine the usefulness of this algorithm in the field of ASR (Kleinschmidt et al., 1998). A significant increase in recognition performance could be observed. However, to evaluate the combination of binaural filtering and the PEMO/LRNN isolated digit recognition system in more realistic conditions, actual recordings were taken as training and test data. The experimental setup is shown in Fig. 8. The speech signals from the ZIFKOM corpus and different noise signals were re-recorded, both in an anechoic chamber and in a moderately reverberant seminar room (average reverberation time of 0.5 s). The same loudspeaker was placed at different azimuth angles 2.5 m away from the Oldenburg dummy head on the horizontal plane. The signals from the built in microphones were directly recorded on hard disk. Later on, speech and noise signals were mixed at different SNRs and azimuthal directions and processed by the binaural algorithm. Finally, the left output channel underwent PEMO feature extraction and LRNN classification.

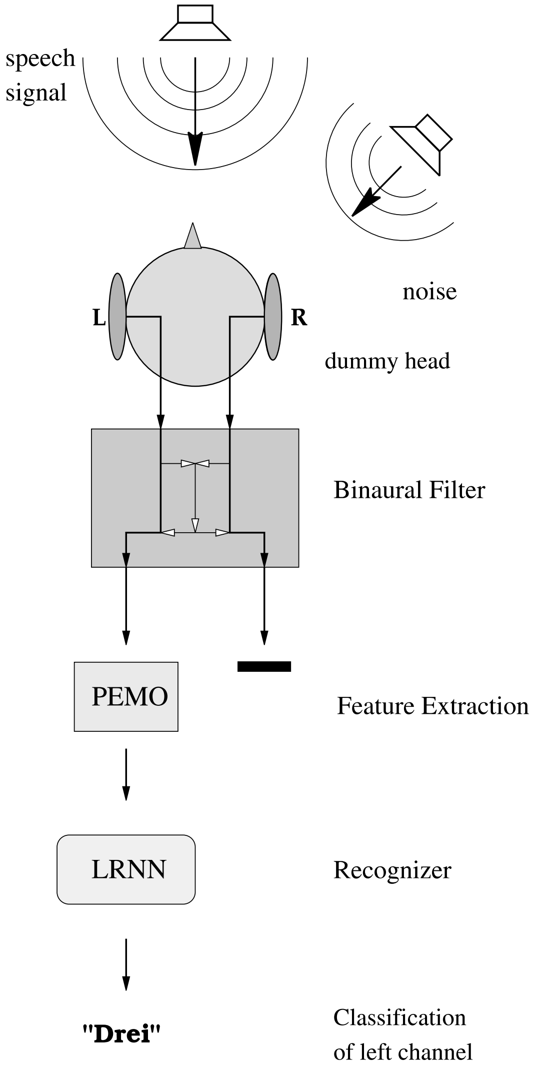

Fig. 8. Setup of binaural isolated digit recognition experiment. Speech signals and noise were recorded separately with the same setup of Oldenburg dummy head and loudspeaker configuration. The signals from different source locations were then mixed, processed by the binaural filter and finally the left channel used for feature extraction and classification.

The SNR was not easy to determine because speech and noise sources were placed at different azimuthal angles relative to the dummy head. Calculating the RMS values at the sources would have meant to ignore the effect of the dummy head related transfer function and required a high technical effort when recording the signals. Instead, the RMS values were calculated using speech and noise arriving from frontal sources at 0° azimuth and the corresponding gain was applied to the lateral noise recordings. In all cases, the speech source was situated in front of the dummy head, while the noise source was located at different angles to the right. The LRNN training was always carried out on clean speech data, which were recorded and filtered the same way as the test corpus, i.e. the reverberant test data were evaluated using an LRNN trained on reverberant training data.

5.2. Results

The speaker-independent recognition results for isolated German digits recorded in the anechoic chamber are shown in Fig. 9. When applying the binaural filtering algorithm the error in CCITT noise (30° and 60°) drops significantly. The effect is less pronounced for a jammer source at 30° azimuth and very low SNR levels. A possible explanation might be the tuning of the directional filter in this set of experiments to no attenuation between 0° and 20° and maximum attenuation for all sources located at directions over 40°. The maximum gain in recognition performance (60°) was about 60% in total at 0 dB SNR, which corresponds to an effective gain in SNR of approximately 10 dB at 90% level. As expected, the directional filter yields no improvement in the 0° case, where speech and noise source are not spatially separated. As no negative effect can be observed either, possible degradations of the speech signal by artifacts of the processing are not ‘noticed’ by the recognition system. Furthermore, in the case of clean test data the error rate has not changed significantly by applying the binaural filtering (see Table 2).

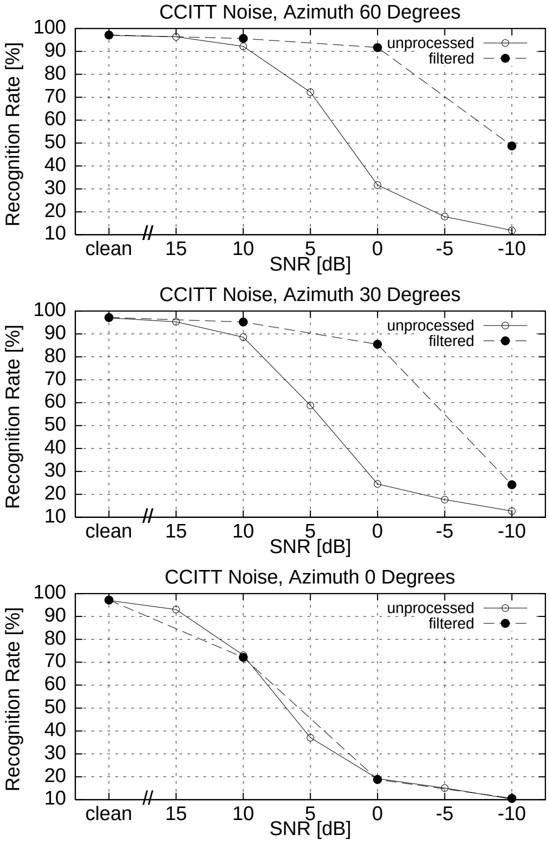

Fig. 9. Speaker-independent, isolated digit recognition rates in CCITT noise as function of SNR in anechoic condition for different angles of noise source location.

Table 2
Speaker-independent isolated word recognition rate in % on clean test data for different reverberative environments with binaural filtering and dereverberation (BF) and no processing (no)

	no	BF
Anechoic chamber	97.3	97.2
Seminar room	96.0	95.6

This is also true for the experiments in reverberant environment (Fig. 10). In this case, the overall performance is worse than in the anechoic chamber. The PEMO/LRNN recognition system suffers from reverberant conditions. Even the use of training data recorded in the same room yields higher error rates for clean test data than in the anechoic case (see Table 2). Moreover, the binaural filter seems to be less effective under reverberant conditions, resulting in a less pronounced enhancement of recognition rates for CCITT noise at 30° and 60° azimuth. This observed effect coincides with the results from speech intelligibility tests (Wittkop et al., 1997), where improvements

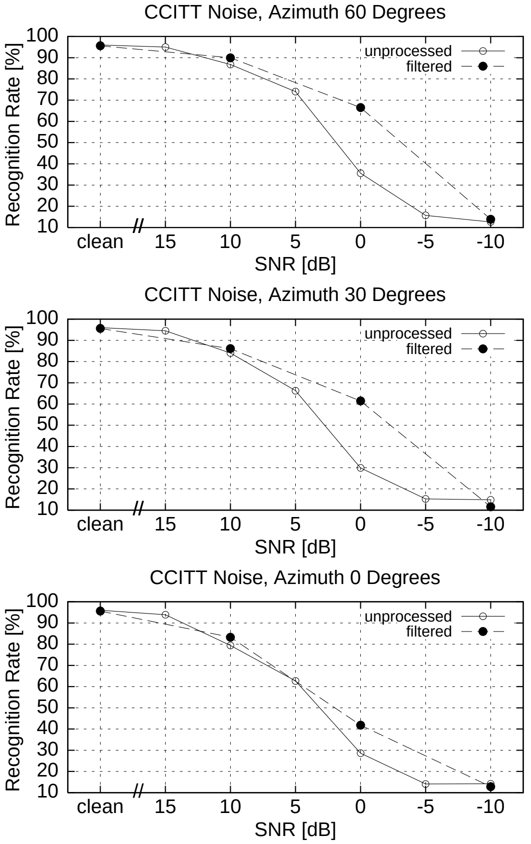

Fig. 10. Speaker-independent, isolated digit recognition rates in CCITT noise as function of SNR in reverberant condition for different angles of noise source location.

on speech perception thresholds were found only in non-reverberant conditions and a small number of jammer sources.

6. Direct comparison of monaural and binaural speech enhancement methods

It can be concluded from the above experiments that monaural and binaural noise reduction algorithms both have the capability to significantly increase the robustness of the PEMO/LRNN isolated digit recognition system. However, a direct comparison is still missing as the experimental setups and especially the SNR calculations were not directly comparable due to the filtering effect of the dummy head. In addition, the evaluated noise signals did not include background speech as interfering sources. Therefore, a third set of experiments was performed and will be described in this section.

6.1. Setup

For the following experiments, the binaural setup (see Section 5 and Fig. 8) has been used. In some cases, the CCITT noise has been replaced by babble noise [4], which was recorded in a cafeteria. For speech enhancement, either the binaural filter or the monaural Ephraim–Malah scheme were applied. The monaural algorithm was only applied to the left channel of the recording. Again, the experiments were carried out in anechoic and in moderately reverberant surroundings using an LRNN trained on clean, unprocessed speech recorded in anechoic or reverberant conditions, respectively.

6.2. Results

The recognition performance of the PEMO/LRNN digit recognition system with binaural filter (BF), Ephraim–Malah (EM) and no processing for the anechoic chamber recordings are presented in Fig. 11. Both algorithms yield a significant improvement in robustness in CCITT noise (top) of about 60% in total at 0 dB SNR or an effective gain of 10 dB SNR at 90% level. The gain of performance by the two algorithms is of similar size, the Ephraim–Malah scheme being slightly superior. In contrast, the binaural filter is far more successful in the suppression of babble noise (bottom) when speech and noise source are spatially separated. In the displayed case of 30° azimuthal angle between speech and noise source, the binaural filter leads to a 30% gain in total at 0 dB SNR compared to 20% with the monaural filter. As expected, the modulated characteristic of babble noise is a bigger problem for the single-channel speech enhancement. Still monaural schemes like Ephraim–Malah's have their advantages, for

[4] NOISEX database, see (Varga et al., 1992).

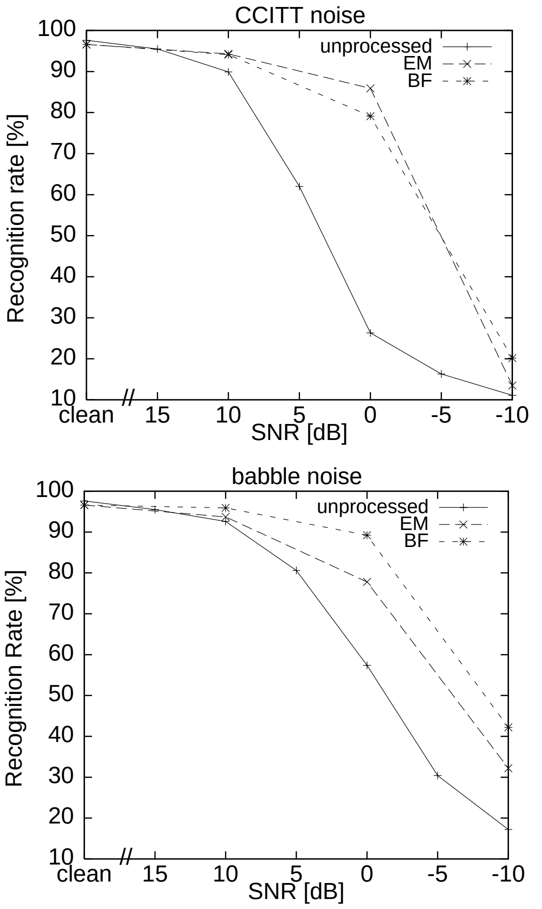

Fig. 11. Speaker-independent, isolated digit recognition rates in CCITT noise (top) and babble noise (bottom) as functions of SNR in anechoic conditions for Ephraim–Malah (EM) and binaural filter (BF) speech enhancement and PEMO/LRNN recognition system. Speech and noise source were separated by a 30° azimuthal difference.

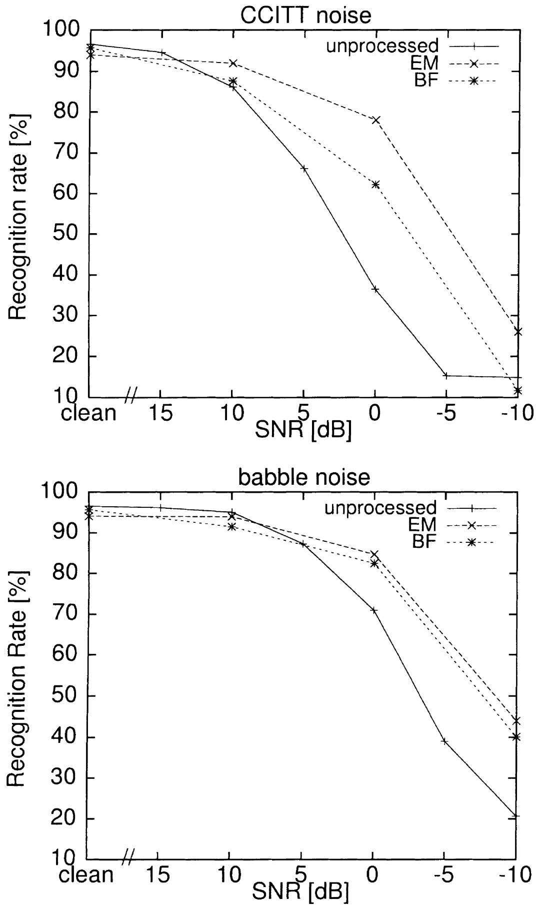

Fig. 12. Speaker-independent, isolated digit recognition rates in CCITT noise (top) and babble noise (bottom) as function of SNR in reverberant conditions for Ephraim–Malah (EM) and binaural filter (BF) speech enhancement and PEMO/LRNN recognition system. Speech and noise source were separated by a 30° azimuthal difference.

example in the case of 0° azimuthal difference between speech and noise source, for which binaural filtering is useless.

The classification results in the moderately reverberant condition are plotted in Fig. 12. As mentioned before, the error rate for clean test data is in all cases higher than in anechoic surroundings. This indicates that the recognition system itself might be disturbed by reverberation, even though the LRNN has been trained on reverberant data. In CCITT noise (top) the monaural and the binaural algorithm yield a significant improvement in recognition performance over the unprocessed alternative. In contrast to the recognition rates obtained in the anechoic chamber, the binaural filter is less effective than the Ephraim–Malah algorithm, scoring a gain of 30% in total at 0 dB SNR compared to 45%. For a quasi-stationary signal like CCITT noise, reverberation is merely a change of spectral characteristics, which is no problem for the Ephraim–Malah algorithm. In contrast to that, the performance of the directional filter and noise source canceler suffers from a higher degree of diffusiveness.

In additive babble noise (bottom), both algorithms yield similar improvement of recognition performance under reverberant conditions, but the gain in recognition rate is much smaller than in anechoic surroundings with 10% in total at 0 dB SNR compared to 20% and 30%. The advantages and disadvantages of monaural and binaural noise reduction schemes, e.g. non-stationarity of the noise signal and reverberation, seem to affect the speech enhancement for ASR systems to a comparable degree. The consecutive application of both algorithms might lead to further synergetic effects and will be evaluated in future studies.

7. Discussion

The auditory model, which was originally developed for predicting human performance in psychoacoustical masking experiments, shows promising results when applied as a front end for ASR. The intention of this quantitative model of auditory processing is to transform an incoming sound waveform into its internal representation. Rather than trying to model each physiological detail of auditory processing, the approach is to focus on the effective signal processing in the auditory system which uses as little physiological assumptions and physical parameters as necessary, while still predicting as many psychoacoustical aspects and effects as possible. A recent study (Tchorz and Kollmeier, 1999) focuses on the amount that each processing stage of PEMO contributes to the robust representation of speech. The results show that the adaptive compression stage is of major importance in this task. The nonlinear adaptation loops yield an enhancement of changes in the input signal and suppression of steady-state portions. When the adaptation stage of PEMO is replaced by a static logarithmic compression of the amplitude in each frequency band (as in common bank-of-filters front ends), the recognition rates in quiet were high but dropped rapidly when the test material was distorted with additive noise (see Fig. 3). Another processing step which contributes to robust recognition is the low pass filter which smoothes the fluctuations in each frequency band after dynamic compression. The filter leads to a band pass characteristic of the amplitude modulation transfer function of the model: slow modulations are suppressed by the adaptation loops, fast modulations by the filter. The maximum in the modulation transfer function in the original model is at approximately 6 Hz. Shifting the maximum to 4 Hz by modified low pass filtering further enhances robustness of ASR in noise. This might be explained by the better correspondence with the average modulation spectrum of speech, which has its maximum at around 4 Hz. Fast fluctuations in the input signal which are not likely to origin from speech are better suppressed with a modified low pass filter. A more detailed study by Kanedera et al. (1999) on modulation processing of ASR front ends supports this hypothesis.

A further improvement of the robustness of ASR systems can be achieved by applying monaural and binaural noise suppression schemes to the disturbed input signal. The PEMO front end has shown to yield robust recognition performance not only against additive noise, but also against possible distortions and artifacts introduced by the noise reduction algorithms. It seems to work especially well when combined with speech enhancement methods originating from digital hearing aid technology. Although (hearing-impaired) human listeners have not been shown to gain a significant advantage from speech enhancement schemes in terms of speech intelligibility, the PEMO/LRNN recognition system obviously benefits to a large degree. This may be caused by two factors: (a) The range of SNRs necessary to obtain 50% intelligibility in normal and most hearing-impaired listeners is still lower than the SNR required to achieve a 50% recognition rate for the ASR systems tested here. Since the performance of noise reduction schemes usually degrade with decreasing SNR, the higher gain in 'intelligibility' for ASR applications might be due to this difference in original SNR level employed. (b) The highly efficient cognitive system of normal and hearing-impaired human listeners is able to compensate for unfavorable SNR conditions by decomposing the incoming sound image into desired speech and undesired background noise. The ease of listening tests and subjective preferences of human subjects indicate that a major cognitive effort is needed for the human brain to make the noise suppression algorithms redundant. This ability is not included in the usual construction of ASR systems. Hence, they have to rely on an appropriate (acoustical) pre-processing to obtain a high enough SNR.

For the purpose of robust ASR, the results are very promising. Besides the major increase of recognition performance which is equivalent to an effective gain in SNR by 5–10 dB in most cases, it is very important that the error rate for clean test data did not change significantly. The positive effect of the noise reduction algorithms was found for all examined types of noise as well as for all SNR and spatial configurations. The exceptions to this general trend were expected, e.g. no increase with binaural filtering at 0° azimuth between speech and noise source, or a limited effect of the Ephraim–Malah algorithm on highly modulated speech noise (babble).

It should be noticed here that in this paper, the ASR systems were always trained on *clean* training data. It cannot be concluded that the advantage of the PEMO front end and especially speech enhancement methods still holds when training is performed differently, e.g. using clean *and noisy* input data. However, since the type and level of disturbing noise in practical conditions is generally not known a priori, it is expected that the advantage of the PEMO front end demonstrated here may still hold in practical applications when untrained noise is encountered.

As a major problem the degraded performance in reverberant environments remains. The PEMO/LRNN recognition system shows no optimal performance even when trained with reverberation, i.e., training on data recorded in the same room as the test data. Also the binaural filter algorithm suffers from a high degree of diffusiveness in the input signals, while the monaural algorithm seems to work similarly well in all surroundings.

The intention of this paper is to demonstrate the usefulness of speech enhancement techniques to already robust PEMO-based ASR systems which had been tested successfully against other front ends in the past. Speech enhancement techniques which were originally designed to increase speech intelligibility of (hearing-impaired) human listeners were combined with auditory-based feature extraction. There is a large variety of other techniques which can partly be applied after feature extraction or are based on a modified classifier. Combining these different approaches to robust speech recognition might yield recognition systems even more robust towards additive and convolutive noise. The idea behind this paper is to show that speech enhancement combined with auditory-based feature extraction might be a promising candidate to play an important role in that task.

8. Outlook

To further evaluate the usefulness of PEMO as front end for ASR systems, experiments with extended vocabulary (more than only 10 digits) or based on sub-word units are necessary. When transformed to PEMO-CEP features, the PEMO internal representation of signals could be examined as a front end for standard HMM phoneme-based recognition systems. In addition, the PEMO/LRNN system needs to be optimized for real-world applications in reverberant environments. This is especially important when it comes to hand-free input devices. Binaural filter algorithms in principal have the capability to reduce reverberation, yet the one applied in this study suffers more by the presence of acoustical echoes than its monaural counterpart. A combination of the binaural filter with a successive Ephraim–Malah noise reduction might result in a further synergistic improvement of performance (see (Meyer and Simmer, 1997) for a combination of binaural and monaural processing). Such a combination has been evaluated already for hearing-impaired subjects wearing hearing aids (Marzinzik et al., 1999).

To aquire the results of speech intelligibility and ease of listening tests, extensive and time-consuming experiments had to be carried out with (hearing-impaired) human subjects. Although the requirements and SNR conditions are somewhat different between human listener speech intelligibility tests and ASR experiments, the methods proposed here suggest an 'objective' way to evaluate noise reduction algorithms also for other types of applications, e.g., hearing aid and telecommunication technology.

Acknowledgements

We are grateful to Mark Marzinzik and Thomas Wittkop for supplying us with their implementation of the speech enhancement algorithms and much valuable advice. Thanks also to Volker Hohmann for fruitful discussions and support. We express our gratitude towards Klaus Kasper and Herbert Reininger, from Institut für angewandte Physik, Universität Frankfurt, for supplying us with their LRNN implementation. We thank Angelika Sievers and Holle Kirchner for translating the abstract of this paper into French. Part of this work was supported by Deutsche Forschungsgemeinschaft (KO 942/12-1).

References

Bitzer, J., Simmer, K., Kammeyer, K.-D., 1999. Multi-microphone reduction techniques for hands-free speech recognition – a comparative study. In: Proceedings of the Workshop on Robust Methods for Speech Recognition. Tampere, Finland, pp. 171–174. .

Blauert, J., 1997. Spatial Hearing, revised ed. MIT Press, Cambridge, MA.

Bodden, M., Anderson, T., 1995. Binaurale automatische Spracherkennung im Störschall. In: Fortschritte der Akustik – DAGA 1995. DEGA, Oldenburg, pp. 1145–1148.

Cappé, O., 1994. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE Trans. Speech Audio Process. 2 (2), 345–349.

Colburn, H.S., 1996. Computational models of binaural processing. In: Hawkins, H.L., McMullen, T.A., Popper, A.N., Fay, R.R. (Eds.), Auditory Computation, Springer Handbook of Auditory Research. Springer, New York, pp. 332–400 (Chapter 8).

Dau, T., Püschel, D., Kohlrausch, A., 1996a. A quantitative model of the 'effective' signal processing in the auditory system: I. Model structure. J. Acoust. Soc. Am. 99 (6), 3615–3622.

Dau, T., Püschel, D., Kohlrausch, A., 1996b. A quantitative model of the 'effective' signal processing in the auditory system: II. Simulations and measurements. J. Acoust. Soc. Am. 99 (6), 3623–3631.

Dau, T., Kollmeier, B., Kohlrausch, A., 1997. Modeling auditory processing of amplitude modulation. I+II. J. Acoust. Soc. Am. 102 (5), 2892–2919.

Derleth, R.P., 1999. Temporal and compressive properties of the normal and impaired auditory system. Ph.D. thesis, Universität Oldenburg.

Durlach, N.I., 1972. Binaural signal detection: Equalization and cancellation theory. In: Tobias, J.V. (Ed.), Foundations of Modern Auditory Theory. Academic Press, New York, Vol. II, Chapter 10, pp. 369–462.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. ASSP-32 (6), 1109–1121.

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean-square error log spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. ASSP-33 (2), 443–445.

Fischer, A., Stahl, V., 1999. On improvement measures for spectral subtraction applied to robust automatic speech recognition in car environments. In: Proceedings of the Workshop on Robust Methods for Speech Recognition. Tampere, Finland, pp. 75–78.

Francis, I.F., Anderson, T.R., 1997. Binaural phoneme recognition using the auditory image model and cross-correlation. In: Proc. ICASSP 97. pp. 1231–1234.

Gelin, P., Junqua, J.-C., 1999. Techniques for robust speech recognition in the car environment. In: Proc. Eurospeech 1999. Budapest, Hungary, Vol. 6, pp. 2483–2486.

Ghitza, O., 1988. Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in noisy environment. J. Phonetics 16, 109–123.

Hansen, M., Kollmeier, B., 1997. Using a quantitative psychoacoustical signal representation for objective speech quality measurement. In: Proc. ICASSP 1997. Munich, pp. 1387–1391.

Hermus, K., Dologlou, I., Wambacq, P., Van Compernolle, D., 1999. Fully adaptive SVD-based noise removal for robust speech recognition. In: Proc. Eurospeech 1999. Budapest, Hungary, Vol. 5, pp. 1951–1954.

Holube, I., Kollmeier, B., 1996. Speech intelligibility prediction in hearing-impaired liseners based on a psychoacoustically motivated perception model. J. Acoust. Soc. Am. 100, 1703–1716.

Jankowski, C., Hoang-Doan, H., Lippmann, R., 1995. A comparison of signal processing front ends for automatic word recognition.. IEEE Trans. Speech Audio Process. 3 (4), 286–293.

Kanedera, N., Arai, T., Hermansky, H., Pavel, M., 1999. On the relative importance of various components of the modulation spectrum for automatic speech recognition. Speech Communication 28, 43–55.

Kasper, K., Reininger, H., 1999. Evaluation of PEMO in robust speech recognition. J. Acoust. Soc. Am. 105 (2), 1175.

Kasper, K., Reininger, H., Wolf, D., Wüst, H., 1995. A speech recognizer with low complexity based on RNN. In: Neural

Networks for Signal Processing V, Proceedings of the IEEE Workshop. Cambridge, MA, pp. 272–281.

Kasper, K., Reininger, H., Wolf, D., 1997. Exploiting the potential of auditory pre-processing for robust speech recognition by locally recurrent neural networks. In: Proc. ICASSP 1997. pp. 1223–1226.

Kermorvant, C., Morris, A., 1999. A comparison of two strategies for ASR in additive noise: Missing data and spectral subtraction. In: Proc. Eurospeech 1999. Budapest, Hungary, Vol. 6,. pp. 2841–2844.

Kiyohara, K., Kaneda, Y., Satoshi, T., Hiroaki, N., Junji, K., 1997. A microphone array system for speech recognition. In: Proc. ICASSP 1997. pp. 215–218.

Kleinschmidt, M., Tchorz, J., Wittkop, T., Hohmann, V., Kollmeier, B., 1998. Robuste Spracherkennung durch binaurale Richtungsfilterung und gehörgerechte Vorverarbeitung. In: Fortschritte der Akustik – DAGA 1998. DEGA, Oldenburg, pp. 396–397.

Kleinschmidt, M., Marzinzik, M., Kollmeier, B., 1999. Combining monaural noise reduction algorithms and perceptive pre-processing for robust speech recognition. In: Dau, T., Hohmann, V., Kollmeier, B. (Eds.), Psychophysics Physiology and Models of Hearing. World Scientific, Singapore, pp. 267–270.

Kollmeier, B., Sotscheck, J., Kammermeier, 1988. Digitalaufnahme eines Reimtests in deutscher Sprache. Audiol. Akustik 27, 24–27.

Kollmeier, B., Peissig, J., Hohmann, V., 1993. Real-time multiband dynamic compression and noise reduction for binaural hearing aids. J. Rehab. Res. Dev. 30, 82–94.

Marzinzik, M., Kollmeier, B., 1999. Developement and evaluation of single-microphone noise reduction algorithms for digital hearing aids. In: Dau, T., Hohmann, V., Kollmeier, B. (Eds.), Psychophysics, Physiology and Models of Hearing. World Scientific, Singapore.

Marzinzik, M., Wittkop, T., Kollmeier, B., 1999. Combination of monaural and binaural noise suppression algorithms and its use for the hearing-impaired. J. Acoust. Soc. Am. 105 (2), 977.

Meyer, J., Simmer, K., 1997. Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtracion. In: Proc. ICASSP 1997, Vol. 2. pp. 1167–1170.

Mine, R., Kobayashi, T., Katsuhiko, S., 1996. Speech recognition in non-stationary noise based on parallel HMMs and spectral subtraction. Syst. Comput. Jpn. 27 (14), 37–44.

Omologo, M., Matassoni, M., Svaizer, P., Giuliani, D., 1997. Microphone array based-speech recognition with different talker-array positions. In: Proc. ICASSP 1997. pp. 227–230.

Patterson, R.D., Nimmo-Smith, J., Holdsworth, J., Rice, P., 1987. An efficient auditory filterbank based on the gammatone function. Paper presented at a meeting of the IOC Speech Group on Auditory Modelling at RSRE.

Peissig, J., 1993. Binaurale Hörgerätestrategien in komplexen Störschallsituationen, Vol. 88 of 17. VDI, Düsseldorf.

Peissig, J., Kollmeier, B., 1997. Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners. J. Acoust. Soc. Am. 101 (3), 1660–1670.

Seneff, S., 1988. A joint synchrony/mean-rate model of auditory speech processing. J. Phonetics 16, 55–76.

Tchorz, J., Kollmeier, B., 1999. A model of auditory perception as front end for automatic speech recognition. J. Acoust. Soc. Am. 106 (4), 2040–2050.

Tchorz, J., Kasper, K., Reininger, H., Kollmeier, B., 1997. On the interplay between auditory-based features and locally recurrent neural networks. In: Proc. Eurospeech 1997. Rhodes, Greece, Vol. 4, pp. 2075–2078.

Varga, A., Steeneken, H., Tomlinson, M., Jones, D., 1992. The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Technical Report, DRA Speech Research Unit, UK, and TNO, The Netherlands.

Vizinho, A., Green, P., Cooke, M., Josifovski, L., 1999. Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: An integrated study. In: Proc. Eurospeech 1999. Budapest, Hungary, Vol. 5, pp. 2407–2410.

Wesselkamp, M., 1994. Messung und Modellierung der Verständlichkeit von Sprache. Ph.D. thesis, Universität Göttingen.

Wilmers, H., Strube, H.-W., 1999. Noise reduction for speech signals by operations on the modulation frequency spectrum. J. Acoust. Soc. Am. 105 (2), 1092.

Wittkop, T., Albani, S., Hohmann, V., Peissig, J., Woods, W., Kollmeier, B., 1997. Speech processing for hearing aids: Noise reduction motivated by models of binaural interaction. Acustica United with Acta Acustica 83 (4), 684–699.

Wittkop, T., Hohmann, V., Kollmeier, B., 1999. Noise reduction strategies employing interaural parameters. J. Acoust. Soc. Am. 105 (2), 977.

Zerbs, C., Dau, T., Kollmeier, B., 1999. Modelling the effective binaural signal processing in detection experiments. In: Dau, T., Hohmann, V., Kollmeier, B. (Eds.), Physophysics, Physiology and Models of Hearing. World Scientific, Singapore, pp. 277–310.