IMPROVED SPEECH RECOGNITION USING HIGH-PASS FILTERING OF SUBBAND ENVELOPES

H.G. Hirsch[1], P. Meyer[2] and H.W. Ruehl[2]

[1] Technical University of Aachen, Templergraben 55, D-5100 Aachen, FRG
[2] Philips Kommunikations Industrie AG, Thurn-und-Taxis-Str. 14, D-8500 Nuernberg, FRG

Abstract

To cope with variability of speech introduced by e.g. background noise, different conditions for training and recognition, and by changes of transmission or speaker characteristics, two series of experiments were carried out using high-pass filtering of spectral envelopes for speech recognition.

In the first series of experiments, FIR high-pass filters were evaluated to reduce the influence of background noise for isolated-word recognition. Introducing high-pass filtering, S/N ratio may be increased by 7 dB while keeping recognition rate constant at 95%.

In the second line of experiments, IIR high-pass filtering was examined to improve speaker independency and robustness of a connected-words recogniser. For 7-digit strings, the digit error rate was reduced from 21.8% to 2.0%.

1. Introduction

The recognition rate of speech recognisers decreases with increasing variability of the input speech to be recognised. One aspect increasing the variability can be varying background noise. E.g., for voice dialling in a car, names may be uttered either at stand-still or while driving, and speech may be captured by a handset or by a hands-free microphone. In each case, the level of background noise and the transmission characteristics are severely affected. Furthermore, the recognition system cannot be trained such that all potential background noise conditions and transmission characteristics are covered during training.

Another aspect specifically valid for speaker independent recognition systems is that typical systems achieve good recognition results on average, but that the error rate increases for specific speakers.

A preprocessing method will be presented improving recognition accuracy for all these conditions. This method, employing only one microphone channel, is based on high-pass filtering of the spectral envelopes in subbands. An advantage of this method is that no speech pause detection is required (e.g. in comparison with the well known noise reduction technique of

The preprocessing method was implemented in two different speaker independent recognisers, an isolated word and a connected-words recogniser. In the case of the isolated word recogniser, high-pass filtering of the magnitude spectral components is done with a FIR filter. In the other case, logarithmically scaled spectral values are filtered with an IIR filter.

2. Recognition Algorithms

Both recognisers used in this evaluation are based on short term spectral analysis.

A special integrated circuit (NEC 7763) is used for the isolated word recogniser. This chip consists of a 16 channel filterbank in the frequency range of up to 6 kHz. Centre frequencies are equally spaced according to the Bark scale. An estimation of the short term subband energies is done every 16 ms. Word boundary detection is based on energy thresholds. The number of spectral vectors within a word is reduced by a factor of about 2 using a trace segmentation algorithm. The speaker independent reference templates in each word class are calculated using a clustering algorithm. Comparison of test and reference templates is done with a dynamic time warping (DTW) algorithm.

To calculate feature vectors, the connected-words recogniser digitally processes speech signals sampled at a rate of 8 Khz. Segments of 256 samples are weighted with a Hamming window and are transformed by an FFT. The power spectral values are smoothed and downsampled to 15 spectral components equally spaced on the Bark scale. The components are scaled logarithmically and normalized to the energy of the speech segment. The energy value is also taken as a 16th component of the feature vector describing one speech segment. Feature vectors are calculated every 12 ms.

The main differences in the signal analysis of the recognisers are, that the isolated word recogniser uses a bandwidth of 6 KHz and linear components, whereas the connected-words recogniser employs only a frequency range up to 4 kHz with regard to telephone applications, and that its spectral components are scaled logarithmically.

The connected word

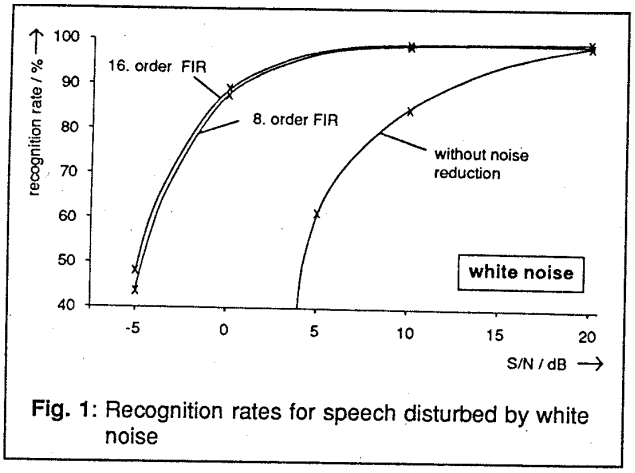

Fig. 1: Recognition rates for speech disturbed by white noise

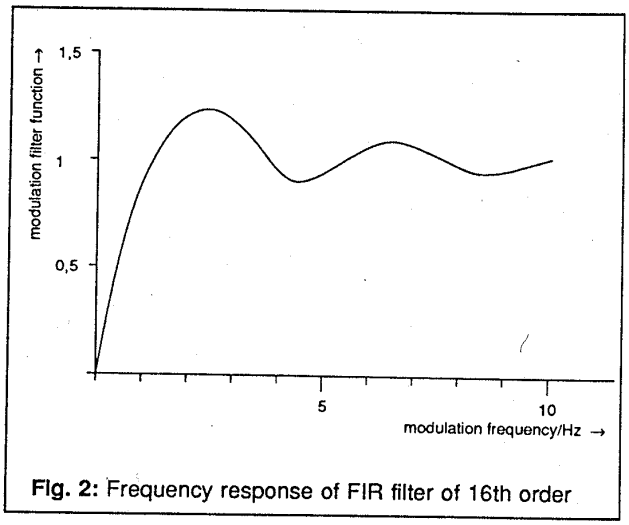

Fig. 2: Frequency response of FIR filter of 16th order

ding training utterances. For recognition, a dynamic time warp algorithm is used, too. A more detailled description of the connected-words recogniser's algorithms may be found in /3/.

3. Improved Word Recognition of Noisy Speech

Earlier studies have shown that speech recognition can be improved for hands-free speech input in reverberant rooms by high-pass filtering the temporal contours of the envelopes in subbands /4/. It is known that the reverberation has an effect similar to low-pass filtering the envelopes within subbands.

This kind of spectral preprocessing by high-pass filtering may be used for noise suppression, too. If the noise is stationary or only slowly changing over time, the contributions to the subband energies will be an almost constant offset. High-pass filtering of the envelopes can reduce these offsets in the subbands. This procedure is closely related to the wellknown spectral subtraction technique. However, it is interesting to note that no speech pause detection is required.

High-pass filtering as preprocessing technique has been integrated into the word recogniser. The short term energy values are high pass filtered using identical FIR filters for each of the 16 spectral components. The results as shown in Fig. 1 lead to an optimal FIR filter order of 16. White Gaussian noise is used to disturb speech in this example. The filter transfer function of this FIR filter of order 16 is shown in Fig. 2.

Tests were done using a vocabulary consisting of 30 German words which might be used to serve a bank automaton. Three reference templates were calculated for each word class. 300 test words of ten speakers were mixed at various S/N-ratios. A considerable improvement was obtained using this noise reduction technique.

Reference templates have been estimated from undisturbed speech signals in the case without noise suppression. In the other case, references were also taken from undisturbed speech, but were preprocessed with the high-pass filtering technique. A gain of about 12 dB can be obtained at a recognition rate of 95 %.

The term modulation frequency is introduced to describe the temporal fluctuation of subband energy. Only modulation frequencies of up to 25 Hz exist in speech signals, with the long-term maximum located at about 3 Hz. The best recognition rates were obtained for the FIR filter as shown in Fig. 1, where only components up to 2 Hz are considerably suppressed. Some further recognition results are shown in Fig. 3.

In this second application, the improvement is not as significant as in the case of disturbance by white noise. The reason for this is that car noise is not stationary due to e.g. changing the gear or using the blink operator. A gain of about 7 dB can be obtained for a recognition rate of 95 %, which is 5 dB less than for white noise, but still a significant gain.

4. Improving Speaker Independence

Based on the results from the tests to improve noise resistance using high-pass filtering, a second set of experiments was set up to evaluate the use of high-pass filtering of modulation frequencies together with a connected-words recogniser to

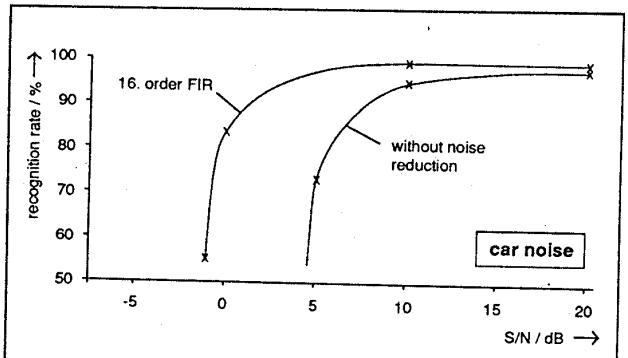

Fig. 3: Recognition rates for speech disturbed by car noise

increase speaker independency and robustness against environment variations.

For this purpose, a corpus was used containing digits and digit strings collected at five different locations in Germany and Austria spoken by 90 male speakers. Both versions of '2', spoken as either 'zwo' or 'zwei' being highly confusable with 'drei' were used.

In order to cover as much variability of speech as possible, collecting procedures and environment conditions were by no means standardized. The collection was done partly via telephone, partly in laboratories and office rooms, partly in an anechoic chamber. Due to differing background noise, SNRs ranged down to 15 dB, with an average SNR higher than 30 dB. This data base called MMIX includes isolated digits from 78 speakers, 3-digit strings from 3 speakers, and 7-digit strings from 12 speakers.

For comparison purposes, a sub-corpus MPFH was constructed, containing only isolated digits, each digit spoken three times in a quiet room by 56 different speakers from the Hamburg area. This homogenous corpus served to control the influence of the inhomogenity of data base MMIX.

Due to availability of digit strings from only few speakers, training was based completely on isolated digits. Only one reference pattern was calculated per digit. Reference patterns MMIXR were created employing 26 speakers from MMIX using 62 to 82 utterances per digit. For MPFH references, 108 utterances from 36 speakers were used per reference. These subsets are sufficient for a speaker independent training, as recognition tests on training sets and independent sub-corpora showed no significant differences in recognition rate.

Using both corpora, recognition tests were run to evaluate the best-behaving FIR filters of the noise reduction experiments, and some IIR filters. It turned out, that in combination with logarithmically scaled feature components, a simple 1st order IIR high-pass with an impulse response

$$y(n) = x(n) - x(n-1) + 0.7 \cdot y(n-1)$$

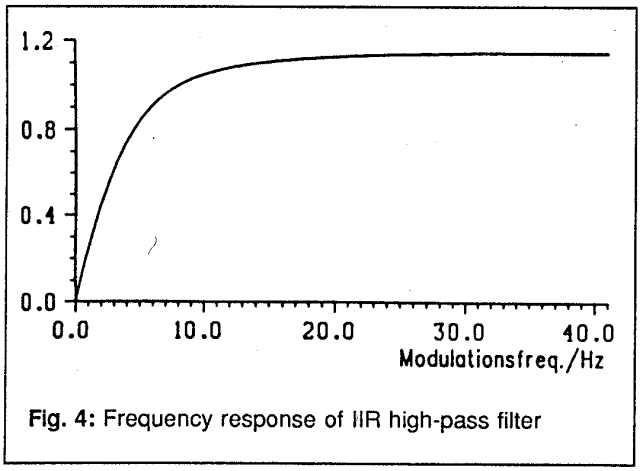

Fig. 4: Frequency response of IIR high-pass filter

corpus / string length	without HP filter				with HP filter			
	subs	del	ins	err/%	sub	del	ins	err/%
MPFH / 1	6	1	19	1.41	7	2	9	0.97
MMIX / 1	366	2	208	19.07	14	13	12	0.96
MMIX / 3	94	0	25	11.53	9	1	0	0.97
MMIX / 7	336	58	226	21.82	23	34	1	2.04

Tab. 1: Digit error rates for speaker independent recognition

performed superior to all FIR filters. The frequency response of the high-pass as given in Fig. 4 shows that the 1st order HP has a cut-off frequency near 4.5 Hz, suppressing significantly more low frequency parts of the modulation frequency than the FIR filter behaving best in noise reduction (see Fig. 2). This makes sense, as it is known that the relevant parts of the modulation frequency spectrum for perception are in the area from 10 to 25 Hz, whereas speaker specific information dominates for frequencies below 10 Hz.

4.1. Speaker Independent Recognition

Using corpus MPFH and references generated from this corpus, high-pass filtering of modulation frequency improves error rates from 1.41% to 0.97% as given in Tab. 1, by halvening the number of insertions. This is only a minor improvement for a homogenous corpus.

For the inhomogenous corpus MMIX, digit error rates of about 1% for isolated digits and 3-digit strings, and 2% for 7-digit strings were achieved with high-pass filtering, which is comparable to the results on corpus MPFH. The major difference is in the results without filtering, as, due to different SNRs, frequency responses etc., between 12% an 22% errors occur.

In the previous experiment, references were created using a subset of MMIX, and, although this corpus is inhomogenous, the reference patterns will try to optimally match the corpus-inherent inhomogenity. But we still do not know how the references copy with variability not observed in the training corpus. For this reason, MMIX references were used to recognise digit strings from an independent corpus MHS. MHS contains utterances spoken by eight males, collected via telephone handset in a running car. Each speaker uttered 44 single digits, 44 3-digit strings, and 100 7-digit-strings. Speaker dependent average signal-to noise ratios were 16 - 25 dB for single digits, 9 - 16 dB for 3-digit strings, and 8 - 13 dB for 7-digit strings.

test corpus / string length	errors / %
MHS / 1	0.6
MHS / 3	3.7
MHS / 7	5.3

Tab. 2: Recognition of noisy speech using speaker independent references MMIX

Results are shown in Tab. 2. For the "clean" single digits with SNRs close to the worst case training patterns, recognition is even better than for MMIX. for digit strings, errors increase to 3.7% resp. 5.3%, both due to decreasing SNR and coarticulation effects. This compares to more than 100% errors for recognition without high-pass filtering, due to the fact that besides frequent substitutions of digits, lots of insertions occured with background noise being mistaken for speech.

4.2 Speaker Dependent Recognition

High-pass filtering of modulation frequency improves recognition with increasing variability of speech. To find out whether some accuracy is lost by high-pass filtering for very homogenous speech, we experimented with some speaker dependent corpora. If the environment is kept constant, then high-pass filtering does not significantly improve speaker dependent recognition. With very high background noise (SNR < 5 dB) both for training and recognition, IIR high-pass filtering even makes recognition worse.

if the environment cannot be controlled, then the IIR high-pass also improves speaker dependent recognition. If for example the corpus MHS is used for speaker dependent training, and a corresponding corpus MHF collected in hands-free mode is used for recognition, then high-pass filtering reduces errors from 148% (due to insertions) to 11.9%.

5. Conclusions

Two very similar preprocessing methods have been presented which are based on high-pass filtering of the envelopes in subbands. A considerable improvement of recognition can be obtained for various conditions of speech input. Both methods are able to reduce background noise without speech pause detection.

The FIR filtering of linearly scaled modulation frequency components seems to be optimal for noise reduction. Estimating the noise resistance for the IIR filter in a similar way as has been done for the FIR filter in Fig. 2, it will probably turn out that the IIR filter curve lies somewhere in the middle between the curves for the FIR and the curve without noise reduction, mostly due to the fact that the logarithmic and not the linear spectral values are filtered.

On the other hand is the broad increase of robustness caused by the fact, that filtering of log. components not only reduces noise influence, but also frequency response of transmission line, microphone, and characteristics of the speaker specifc long term spectrum.

The question whether a combination of two high-pass filters, one filtering linear components to improve noise resistance, a second one filtering after having taken the logarithm to improve robustness, will combine the advantages of both filtering approaches will be approached in some future experiments.

References:

/1/ P. Vary: Noise Suppression by Spectral Magnitude Estimation - Mechanism and Theoretical Limits, Signal Processing, 1985, pp. 387-400

/2/ H.G. Hirsch, H.W. Ruehl: Automatic Speech Recognition in a Noisy Environment, Proc. EUROSPEECH, 1989, pp. 652-655

/3/ H.W. Ruehl et al.: Speech Recognition in the Noisy Car Environment, Speech Communication, vol. 10, no.1, 1991, pp. 11-22

/4/ H.G. Hirsch: Automatic Speech Recognition in Rooms, Proc. EUSIPCO, 1988, pp. 1177-1180