Forward Masking for Increased Robustness in Automatic Speech Recognition

Sascha Wendt, Gernot A. Fink, Franz Kummert

Faculty of Technology, Bielefeld University,
P.O. Box 100131, 33501 Bielefeld, Germany
{swendt,gernot,franz}@techfak.uni-bielefeld.de

Abstract

In automatic speech recognition mel-frequency cepstral coefficients (MFCC) or linear predictive cepstral coefficients (LPCC) are features commonly used today. However, their calculation considers only a few features of the auditory system. On the assumption that the human representation of speech is an optimal representation, considering more features of the auditory system might lead to a better performance of automatic speech recognition systems. In this paper a model proposed by Strope and Alwan [1], which relies on the human acoustic perception and allows to consider the effect of forward masking, is incorporated after some modifications into an automatic speech recognition system with a MFCC-based front-end. The extended system is evaluated on recognition tasks, that are closer to real recognition than (connected) digit recognition commonly used in the literature. The evaluations show an increased robustness of the speech recognition system with forward masking on all recognition tasks, but especially on data recorded in noisy environments.

1. Introduction

Despite the many research efforts the performance of current automatic speech recognition systems is still not as good as the human speech recognition. One reason is the human acoustic perception. There are several features of the auditory system, which are important for the human speech perception but are not considered in automatic speech recognition (ASR). If some of these features would be considered, the performance of ASR systems could possibly be increased. One example for this assumption is the auditory based distortion of the frequency spectrum with a mel-filter bank, which has been established in most ASR systems.

The different approaches to realize some of the auditory features in ASR systems can be roughly subdivided into two groups: On the one hand ASR systems are developed, which feature extraction nearly completely relies on auditory models (e.g. [2, 3]). On the other hand there are different approaches to improve established feature extraction methods like mel-frequency cepstral coefficients (MFCC) or linear predictive cepstral coefficients (LPCC) with auditory features like forward masking (e.g. [1, 4]).

This paper follows the second approach. A model for forward masking, that was proposed in [1], will be described in section 3. This model is incorporated after some modifications, which are described in section 4, into an ASR system with a MFCC-based front-end. The model provides a simple method to calculate the effect of forward masking and can be incorporated without major modifications of the ASR system.

Most publications on ASR systems with auditory features are only evaluated on clean speech with added artificial noise. Additionally, often only small vocabularies like digits are used. In this paper, however, the extended ASR system is evaluated on speech data, which was recorded under real noisy conditions in a car. Additionally evaluations on a conversational speech and on a continuous speech recognition task with 5000 word vocabularies are carried out.

2. Forward masking

Forward masking effects a decreased audibility of a stimulus if it follows a previous stimulus. This effect is shown in figure 1:

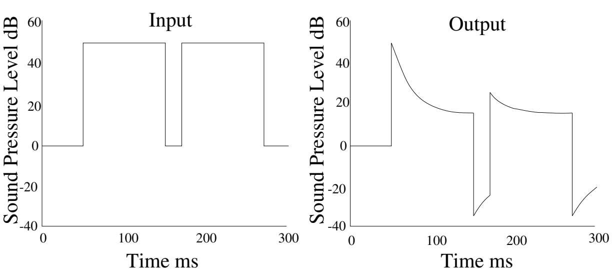

Figure 1: *Illustration of the forward masking effect: a long stimulus masks a second stimulus (following [1].*

The figure shows two consecutive stimuli. The first stimulus (masker) is adapted by the auditory system. This effects a decreasing sensation of the stimulus. The following stimulus (probe), which is also adapted, is additionally masked by the previous stimulus. This forward masking effect leads to a lower sensation at the beginning of the second stimulus compared to the first stimulus, which was presented without a previous stimulus.

The effect of forward masking depends on several conditions. First it depends on the time difference between masker and probe: with an increased delay of a probe the effect decreases. The effect also depends on the time duration of the masker: with an increased duration a following probe is less audible. On the other hand very short stimuli like impulses effect no forward masking. The intensity of the masker is important, too: with an increased intensity the effect increases as well. Because the effect additionally depends on the frequency of masker and probe, for every frequency band used in the ASR system a set of parameters is necessary to represent these dependencies.

3. Related work

In current research on speech recognition there are several approaches to consider the effect of forward masking in ASR systems.

In Proc. European Conf. on Speech Communication and Technology, volume 1, pages 615-618, Aalborg, Daenemark, 2001.

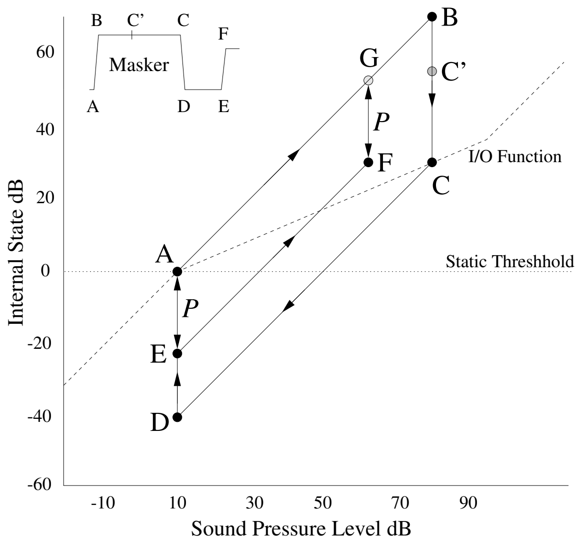

Figure 2: *Output trajectory of the model within the I/O graph with the important points A-F (following [1]). The corresponding observed signal is shown in the upper left.*

In [2] and [5] for example feedback loops are used. Each of these loops has its own time constant and is charged by a signal respectively discharged, if there is no signal. The charging state of these loops then determines the amount of forward masking of the current stimulus. This model was incorporated into an ASR system, which feature extraction nearly completely relies on auditory features.

In [4] the effect of forward masking is considered by a weighted sum of previous signals, which is then subtracted from the current signal. This model is incorporated into an ASR system, which uses a modified MFCC feature extraction.

Another model is the dynamic adaptation presented in [1]. A similar model was used in [6]. This model is described very well and with all necessary details. It relies on the human acoustic perception and can be incorporated without major modifications of the existing ASR system. For this reasons this model was chosen to consider the effect of forward masking and, therefore, will be described in more detail.

For this model an input/output (I/O) function is specified. This I/O function is motivated by the response of the basilar membrane [7] and provides the adaptation target. After every input an adaptation mechanism adjusts an internal state towards this adaptation target. The target corresponds to the human acoustical perception of a stimulus, that is presented over a longer time and not masked by a previous stimulus. The difference between the dynamic internal state and the static target allows to calculate the effect of forward masking as described next.

Figure 2 shows the I/O function within the I/O graph of the model. In the following the trajectory of the internal state is described. The signal observed in the description is shown besides in the upper left of the figure: At point A the masker is switched on and at point D completely switched off. At point E the probe is switched on.

At the beginning of the signal the internal state is positioned at point A: The auditory system, that motivates the I/O function as described above, perceives no signal and there is no masking effect caused by a previous signal. After the onset of the masker, the output trajectory moves instantly along a diagonal to point B, because the output initially tracks abrupt changes of the input. While the masker lasts for a longer time, the model adapts the masker by moving the internal state towards the target C on the I/O function. If the masker lasts long enough, the internal state reaches the target. After the masker offset, the trajectory falls downwards along a diagonal to point D. At this point the distance between the diagonal and the I/O function is constant and the internal state moves slowly towards the target A. This movement represents the decreasing forward masking effect with an increasing probe delay.

Still before the internal state reaches the target A, the probe is switched on at point E. This onset causes once more an abrupt diagonal movement of the internal state to point F. This point is below the point G, that would have been reached by a probe onset without a previous masker. Therefore, the model has masked the probe. The amount of masking (P) corresponds to the difference between the reached point E before probe onset and the target A on the I/O function.

The duration of a masker can also be considered in this model: if the duration of the masker in figure 2 decreases, the adaptation mechanism doesn't reach the target on the I/O function. That means, that the masker can't be completely adapted by the model. Accordingly, after the masker offset the internal state is positioned above the target C (e.g. at point C') and all following points are shifted up by the distance between the target C and the reached point C'. Therefore, the amount of masking (P) is decreased. This means, that the model considers a smaller masking effect caused by a decreased masker duration.

All approaches to consider forward masking in ASR systems mentioned in this section were evaluated on (connected) digit recognition tasks and have shown a better performance, especially on noisy speech data. On the other hand, most of the models lead to a worse performance on clean speech data. For a more thorough evaluation of the ASR systems performance, however, evaluations on recognition tasks are necessary, that are closer to real recognition tasks than simple digit recognition.

4. Adjusting the model

For the application of the described model with the values for the necessary parameters taken from [1], the energies of the frequency groups must be converted into Decibel. However, the required reference sound intensity level for this conversion is unknown, because it depends on the input channel. For this reason a dynamic energy histogram is calculated on the speech data. This histogram is used to normalize the total energy maximum to 60 dB. The offset obtained for this normalization is then utilized as reference sound intensity level.

The value of 60 dB for the normalization was chosen, because it corresponds to the approximated minimal sound pressure level of conversational speech ([8], p. 31). A higher value for the normalization could possibly lead to an overestimated masking effect, if the recorded speech was really not above 60 dB. Although the masking effect is underestimated, if the recorded speech was really above 60 dB, an ASR system, which underestimates the masking effect should still be better than an ASR system, that doesn't consider this effect or overestimates it.

After some initial experiments two modifications of the models application were made.

The first modification refers to the adaptation of a masker: this adaptation leads to a partial masking of long stimuli, that are for example caused by a vowel or consonant. In this way the distinction of different long stimuli decreases and, therefore, the adaptation leads to a loss of speech characteristic. In addition strong background noises are adapted. Although this partially masks the background noise, speech signals appearing during this background noise are partially masked, too. Thereby, the similarity of speech signals with different background noises decreases. The second problem was solved in [1] by training two models, one for clean and one for noisy speech data. For the recognition both models were used, and the model with the higher probability determined the word recognized.

This approach has some disadvantages, however. First, if more then one model is used, the required resources for the computations increase. Secondly, for every model a training-set with a defined signal-to-noise ratio is required. This is no problem, if clean speech with and without added artificial noise is used. But if the speech data was recorded under real noisy conditions, this data can't be simply subdivided into sub-sets with clean and with noisy speech or in sub-sets with defined signal-to-noise ratios.

Because of the above-mentioned reasons the applied model doesn't return the internal state. Instead only the maskers calculated by the model and the static threshold are subtracted from the current signal pressure level. Thus the effect of forward masking but not the adaptation is considered in the further feature extraction.

The second modification refers to the static thresholds: the described model only considers signal pressure levels above these thresholds. This leads to a smaller dynamic range of the speech data. Therefore the static thresholds are not considered after calculating the masker effect. This is achieved by adding the static thresholds to the internal state. But now the values calculated by the model can be very low, especially in pauses between speech signals. For this reason minima are introduced. These minima are dynamically established for every frequency band by saving the lowest appearing value before masking. If a value calculated by the model drops under this minimum, the value is restricted to the corresponding minimum.

The described modifications of the models application lead to a simple output of the model: only the maskers are subtracted from the current signal. For the further feature extraction the static thresholds and the adaptation of the current signal are not taken into account.

5. Evaluation

The automatic speech recognition system extended by the forward masking model was tested on different recognition tasks.

The first task contains the acoustic control of different systems in cars, e.g. radio, car-phone or navigation system. The speech data is from the SLACC corpus (**S**poken **La**nguage **C**ar **C**ontrol), which was recorded in Bielefeld, Germany, in different car types under different speed and weather conditions with two microphones: a close-talking microphone (micro 1) and a microphone fixed to the right jamb of the windshield (micro 2), whose recorded data has a worse quality than the data from the close-talking microphone, because it receives many other signals besides the speech signal. The data includes 22 speakers (co-drivers) providing 10984 utterances (read sentences), which was split for the evaluation: the data of 18 speakers with 9243 utterances was used for the training, the remaining 4 speakers with 1741 utterances constituted the test-set. The vocabulary for test and training contained 658 different words.

Additionally, experiments on the German *Verbmobil* appointment scheduling task [9] and on the *Wall-Street-Journal* task (WSJ0) [10] were carried out. For both tasks the data was recorded in clean acoustic environments and therefore contains clean speech. For Verbmobil a 5336 word recognition system was trained on over 32 hours of spontaneous speech and then evaluated on the independent test-set of approximately 35 minutes of spontaneous speech. The training-set used for Wall-Street-Journal (the phonotypical transcription of the vocabulary was supplied by "Carnegie Mellon Pronouncing Dictionary" Version 0.6) includes 84 speakers providing about 15 hours of speech (read sentences from the *Wall Street Journal*). On this data a 4986 word recognition system was trained and then evaluated on approximately 40 minutes of speech provided by 8 speakers.

The feature extraction of the speech recognition system calculates every 10 ms 12 mel-frequency cepstral coefficients with a dynamic mean normalization on the logarithmic output of the mel-filters. The offsets for this mean normalization are calculated before masking but applied after masking. If the mean normalization is not realized as described, there are negative interactions between mean normalization and the forward masking model, which is also applied on the logarithmic output of the mel-filters.

Besides these cepstral coefficients one energy term normalized as described in section 4, together with the smoothed first and second order derivatives of these 13 features are calculated. On this 39 dimensional feature space semi-continuous HMM systems with linear structure, variable number of states and triphone sub-word units were trained. For the evaluations of the Verbmobil and the Wall-Street-Journal test-sets additionally bigram language models (perplexity: 64.4 on Verbmobil, 109.6 on WSJ0) were applied. For more details about the speech recognition system see [11].

Features	SLACC micro 2	SLACC micro 1	WSJ0	Verb-mobil
MFCC	25.8%	17.5%	12.9%	21.8%
masked MFCC	23.6%	16.5%	12.1%	21.6%

Table 1: *Word error rate of the system with and without forward masking on different recognition tasks.*

Table 1 compares the word error rates (WER) of the speech recognition system with and without forward masking on the different recognition tasks. The results show a significant reduction of the WER on the SLACC corpus, where the reduction on the data of the second microphone (relative improvement 8.5%) is more distinct than on the data of the first (close-talking) microphone (relative improvement 5.7%). On Wall-Street-Journal (WSJ0) the reduction of the WER is also just significant (relative improvement 6.2%), on the Verbmobil recognition task it is not. The application of the forward masking model therefore leads to a reduction of the WER not only but especially on speech data recorded under bad or noisy conditions.

Features	SLACC micro 2	SLACC micro 1	WSJ0	Verb-mobil
MFCC	88.3	81.4	10.1	7.2
masked MFCC	72.8	66.2	6.2	5.6

Table 2: *Variance of the word error rate of the system with and without forward masking within the different recognition tasks.*

Besides the reduction of the WER the application of the

	Test-set SLACC (micro 1)						
	Speaker 1		Speaker 2	Speaker 3		Speaker 4	
Features	Set 1 (1286)	Set 2 (1425)	Set 1 (1065)	Set 1 (1149)	Set 2 (1207)	Set 1 (1568)	Set 2 (1767)
MFCC	28.95%	21.08%	25.09%	21.32%	23.09%	5.26%	6.01%
masked MFCC	26.02%	19.90%	23.89%	22.27%	20.13%	5.45%	6.12%

Table 3: *Word error rate of the system with and without forward masking for the different speaker test-sets. The numbers in brackets show the quantity of words in the sub-sets.*

model shows another positive effect: the variance of the WER within the test-sets decreases. This is shown in table 2.

For the calculation of the variance the test-sets were subdivided. On the SLACC corpus every sub-set includes only one recording session of an individual speaker. On Wall-Street-Journal each of the eight sub-sets includes also only one speaker. On Verbmobil the test-set was subdivided into the four different recording locations. Every location test-set was subdivided once more into two sets: the first sub-set includes the utterances of all first speakers, the second subset the utterances of all second speakers of a dialog. This subdivided test-sets were separately evaluated. For the calculation of the variance the WER of every sub-set was weighted with the number of words in this sub-set. For the data of the SLACC corpus recorded with the close-talking microphone (micro 1) this more detailed evaluation is shown in table 3. The variances on the Verbmobil and the WSJ0 test-sets in table 2 are much smaller compared to the SLACC test-sets, because these test-sets include more speakers and the differences between the WER of the best and the worst sub-sets are also much smaller.

The results in table 2 show, that the variance within the test-sets is reduced on all recognition tasks. Therefore, it can be concluded, that the application of the model leads to an increased robustness of the speech recognition system over the recording conditions (speaker, microphone, background) not only in noisy but also in clean environments.

6. Conclusion

In this paper, a model of forward masking was incorporated after some modifications into an ASR system with a MFCC-based front-end. The extended ASR system was evaluated on speech recognition tasks, that are closer to real recognition tasks than (connected) digit recognition commonly used in the literature, and has shown an increased robustness independent of the recording conditions. The variance of the word error rate is reduced on all recognition tasks, and especially on data recorded in noisy environments the application of the model additionally leads to a significant reduction of the word error rate. Therefore, the ASR system with the modified forward masking model can be applied on all recognition tasks and has a better or at least equal performance compared to a system without forward masking.

7. Acknowledgement

This research was supported by the German Research Foundation (DFG) within SFB 360 ”Situated Artificial Communicators”.

8. References

[1] B. Strope and A. Alwan, ”A model of dynamic auditory perception and its application to robust word recognition”, IEEE Trans. on Speech and Audio Processing, 5(5):451–464, 1997.

[2] J. Tchorz and B. Kollmeier, ”A model of auditory perception as front end for automatic speech recognition”, J. Acoust. Soc. Am., 106(4):2040-2050, 1999.

[3] C. H. Yim, O. C. Au, W. Wan, C. L. Keung and C. C. Fung, ”Auditory spectrum based features (ASBF) for robust speech recognition”, Proc. Int. Conf. on Spoken Language Processing, II:979–982, Bejing, 2000.

[4] Y. Ito, H. Matsumoto and K. Yamamoto, ”Forward masking on a generalized logarithmic scale for robust speech recognition”, Proc. Int. Conf. on Spoken Language Processing, III:530–533, Bejing, 2000.

[5] T. Dau, D. Püschel and A. Kohlrausch, ”A quantitative model of the effective signal processing in the auditory system. I: Model structure”, J. Acoust. Soc. Am., 99(6):3615–3622, 1996.

[6] Z. Liu, X. Wu, B. Zhen and H. Chi, ”An auditory feature extraction method based on forward-masking and its application in robust speaker identification and speech recognition”, Proc. Int. Conf. on Spoken Language Processing, II:266–269, Bejing, 2000.

[7] B. M. Johnstone, R. Patuzzi and G. K. Yates, ”Basilar membrane measurements and the travelling wave”, Hearing Research, 22:147–153, 1986.

[8] E. Zwicker and H. Fastl, “Psychoacoustics”, Springer-Verlag, Berlin Heidelberg, 1999.

[9] K. Kohler, G. Lex, M. Pätzold, M. Scheffers, A. Simpson and W. Thon, “Handbuch zur Datenaufnahme und Transliteration in TP 14 von VERBMOBIL – 3.0”, Institut für Phonetik und digitale Sprachverarbeitung, Universität Kiel, 1994.

[10] D. B. Paul and J. M. Baker, “The design for the Wall Street Journal-based CSR corpus”, Speech and Natural Language Workshop, Morgan Kaufmann, 1992.

[11] G. A. Fink, ”Developing HMM-based Recognizers with ESMERALDA”, in V. Matoušek, P. Mautner, J. Ocelíková and P. Sojka, editors, ”Lecture Notes in Artificial Intelligence”, 1692:229–234, Springer, Berlin/Heidelberg, 1999.