EFFECTIVENESS OF PHASE-CORRECTED RASTA FOR CONTINUOUS SPEECH RECOGNITION

Johan de Veth & Louis Boves

A^{2}RT, Department of Language and Speech,
University of Nijmegen,
P.O. Box 9103, 6500 HD Nijmegen, THE NETHERLANDS

ABSTRACT

Phase-corrected RASTA is a new technique for channel normalization that consists of classical RASTA filtering followed by a phase correction operation. In this manner, the channel bias is as effectively removed as with classical RASTA, without introducing a left context dependency. The performance of the phase-corrected RASTA channel normalization technique was evaluated for a continuous speech recognition task. Using context-independent hidden Markov models we found that phase-corrected RASTA reduces the best-sentence word error rate (WER) by 23% compared to classical RASTA. For context-dependent models phase-corrected RASTA reduces WER by 15% compared to classical RASTA.

1. INTRODUCTION

In order to reduce the linear filtering effect of communication channels, different channel normalization (CN) techniques have been proposed (e.g. [1, 2, 3]). Recently, a new, extended version of the classical RASTA filtering technique was proposed and tested in the context of connected digit recognition over the telephone [4, 5, 6]. The results of these connected digit string experiments showed that the recognition performance of phase-corrected RASTA (pcR) is equivalent to the performance of cepstrum mean subtraction (CMS). In addition, it was concluded that the new CN method is better suited for context-independent modeling than classical RASTA (clR), because it removes the left context dependency introduced by clR.

The connected digit string experiments suffered from an important limitation. In the digit vocabulary the average number of different contexts for each phone is small. Therefore, the impact of introducing a left context dependency by using clR is limited. This explains that clR is still capable of outperforming applying no CN and the gain as a result of switching from clR to pcR is small when using context-independent models. This could also explain why we did not find significant differences between the different CN methods that we studied when we used context-dependent models. [6]. Enlarging the test set would reduce the confidence regions such that possible differences could yet become visible in the case of context-dependent models. However, staying in the connected digit domain could never have taken away the limitation due to the small number of different contexts for each phone.

In this paper, we report on experiments using phase-corrected RASTA for a continuous speech recognition task. In this task the average number of different contexts for each phone is much higher. In addition, the amount of training and testing data we used is much larger. For these reasons, the new task is better suited to test the effectiveness of different CN techniques relative to each other. Especially the effects of using context-dependent vs. context-independent models can be well established with the new set-up. Thus, this CSR task will provide a thorough check on our original claims about phase-corrected RASTA [6].

This paper is organised as follows. The telephone database that we used for our experiments is described in section 2. In section 3, the signal processing for our experiments is described. The topology of the hidden Markov models (HMMs), the way these were trained and the recognition task are described in section 4. The results of our recognition experiments are discussed in section 5. Finally, in section 6 we sum up the main conclusions.

2. DATABASE

The speech material for these experiments was collected with an on-line version of a spoken dialogue system which provides public transport information in the Netherlands. This system is an adaptation of a German prototype developed by Philips Research Labs [7, 8]. Speakers were recorded over the public switched telephone network in the Netherlands. Speakers, handset and channel characteristics are not known.

A total of 33,471 utterances was collected. For training we reserved 25,104 utterances (83,876 words corresponding to 8.9 hours of speech excluding leading, intermediate and trailing silent portions of the recordings). The remaining 8,358 utterances (28,048 words corresponding to 3.0 hours speech) were set apart as an independent test set. None of the utterances used for training or test had a high background noise level.

The average number of words per utterance is 3.3; this is rather low, especially when it is compared to ATIS, Wall Street Journal or North American Business News. The short utterances are quite normal in real dialogues between callers and operators in information services. The language of the corpus is Dutch; the speech was spontaneous and unprepared.

3. SIGNAL PROCESSING

Speech signals are in A-law format. After conversion to a linear scale, preemphasis with factor 0.98 was applied. A 25 ms Hamming window that was shifted with 10 ms steps was used to calcu-

late 24 filterband energy values for each frame. The 24 triangular shaped filters were uniformly distributed on a mel-frequency scale (covering 0 - 2143.6 mel). Finally, 12 mel-frequency cepstral coefficients (MFCC's) were derived [9]. In addition to the twelve MFCC's we also computed before CN was applied the twelve first time-derivatives (delta-MFCC's), log-energy (logE) and its first time-derivative (delta-logE). In this manner we obtained 26-dimensional feature vectors.

We applied CN only to the twelve MFCC coordinates of the feature vector. We kept the original values of delta-MFCC's, logE and delta-logE. For CMS the vector of average cepstral coefficients was calculated over the whole utterance (i.e., including leading, intermediate and trailing silent portions of the recorded signal). We used clR with integration factor -0.94 [2]. For pcR we used the same integration factor in combination with a phase-correction filter [4, 5, 6]. During the time-reversal operations required for the phase-correction we used the whole utterance [6].

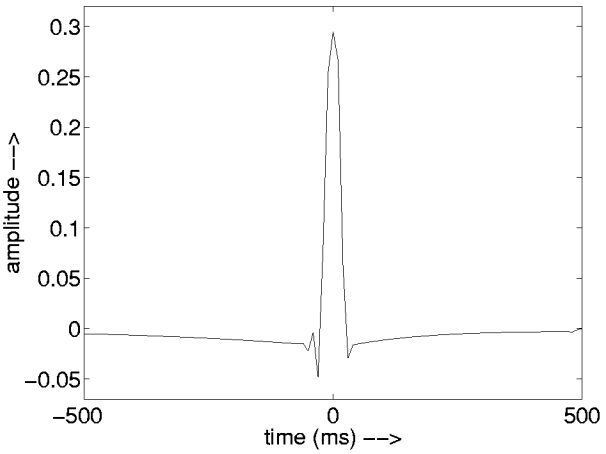

Figure 1: Impulse response of the phase-corrected RASTA filter.

The impulse response of the pcR filter is shown in Figure 1. Due to the zero-phase pcR filter characteristic, the impulse response is essentially symmetric. Symmetry of the impulse response was reported to be one of the key features of optimally designed filters that were calculated in a data-driven approach based on linear discriminant analysis [10].

4. MODELS

4.1. Training context-independent models

37 context-independent phone models were trained. In addition, we used one model for all sorts of noise and one model to describe silence. The phone models and the noise model consisted of six hidden Markov states, where states 2, 4 and 6 shared the emission probability density function with states 1, 3 and 5, respectively. A single-state HMM was used for the silence model. All HMMs were left-to-right where only self-loops, transitions to the next state or to the next state plus one were allowed. The emission probability density functions were described as a continuous mixture of 26-dimensional Gaussian probability density functions (diagonal covariance matrices). In order to be able to study the recognition performance as a function of acoustic resolution, we used mixtures which contained 4, 8, 16 and 32 Gaussians. With a total of 115 HMM states we arrived at CI-HMM systems containing a total of 460, 920, 1840 and 3680 Gaussian densities respectively.

The training lexicon contained 1415 words. Models were initialised using a linear segmentation within the speech portions of the signal, as determined with a silence-speech detector. After initialisation a fixed number of Viterbi optimisation passes was used to further train the models. As a next step the number of Gaussians per state was doubled. To this aim a K-means clustering algorithm was applied using the segmentations obtained in the previous Viterbi pass [7]. After splitting, again Viterbi optimisation was applied. This process of successive splitting and subsequent Viterbi optimisation was repeated until we obtained models with 32 Gaussians per state.

4.2. Training context-dependent models

In order to define context-dependent HMMs we determined all different contexts for each phone in our training material and used a state-tying mechanism to avoid the risk of undertraining. To this aim each phone in our database was considered to consist of three segments, where the first segment corresponded to the first two HMM states, the second segment to states 3 and 4, and the last segment to states 5 and 6. For clustering segments it was assumed that the first segment only depends on the phone immediately to the left of the phone under consideration, the middle segment is independent of the context and the last segment only depends on the phone immediately right to the phone under consideration. During clustering word boundaries were regarded as a special phone. As a consequence, we did not model cross-word context. The number of independent CD phone units to train models for was determined by specifying the minimum number of observations of a phone in a particular left or right context.

In a set of tuning experiments we determined the optimum number of CD phone units for the training database described above. Of course, the same lexicon was used as for the contect independent models. We found that the recognition performance was not critically sensitive to the number of CD phone units. All data in this paper are based on a system with 388 CD phone units. This choice allows us to compare CI-HMMs and CD-HMMs with approximately equal numbers of Gaussian densities. We trained CD-HMMs with 1, 2, 4 and 8 Gaussians per state. In this manner we arrived at CD-HMM systems with a total of 388, 776, 1552 and 3104 Gaussian densities respectively.

For each CD-HMM system we jointly optimized the word entrance penalty and the language model factor. In order to avoid optimization on the actual test set, we used a jack-knifing procedure with the number of sub-sets $N = 4$. We wanted to be able to compare results for the CD-HMM systems to those obtained for the CI-HMM systems. For this reason we used exactly the same division into sub-sets in both cases when we evaluated the recognition performance. In the case of the CI-HMMs we did not use the jack-knifing procedure to determine the optimal values of the word entrance penalty and the language model factor. The optimal values for the CI-HMMs were copied from a previous version of the CSR [8].

4.3. Recognition

The recognition lexicon contained 983 words. 1.2% of the words in the test set were out-of-vocabulary. During recognition the acoustic models were combined with unigram and bigram language models derived from the training data. The test set perplexity of the recognition task was 36.7. For our evaluations we restricted ourselves to the single best recognized sentence. The best-sentence word error rate (WER) was defined as

$$WER = \frac{S + D + I}{N} \times 100\%, \quad (1)$$

where N is the total number of words in the test set, S denotes the total number of substitution errors, D the total number of deletion errors and I the total number of insertion errors. The WER values presented in this paper were obtained by averaging over the WER values obtained for those test utterances that were not used to find the optimal values for the word entrance penalty and the language model factor.

5. RESULTS AND DISCUSSION

5.1. Results for CI-HMMs

We trained and tested CI-HMMs for four different conditions: no channel normalization (NCN), clR, CMS over the whole utterance and pcR over the whole utterance. The WER is shown in Figure 2 as a function of the total number of Gaussians used. Figure 2 shows that clR deteriorates recognition performance compared to NCN. Apparently, removing the channel bias by using clR at the same time introduces so much left context dependency that the potential CN gain is completely annihilated. The results for pcR indicate that the poor performance of classical RASTA is a direct consequence of the phase distortion introduced. By removing the phase distortion the recognition performance is significantly and substantially improved compared to clR. At the highest total number of Gaussians in our CI-HMM approach the WER is reduced by 23% relative to clR. Furthermore, pcR recognition performance is significantly better compared to NCN (except at the suboptimal models corresponding to 4 Gaussians per state). Finally, it can be seen that CMS appears to be preferred over pcR for CI-HMMs corresponding to 4 and 8 Gaussians per state. However, for more complex acoustic models the performance differences become insignificant and pcR performs as well as CMS.

The results shown in Figure 2 are in good agreement with the results we reported earlier in the context of a connected digit recognition task [5, 6]. In that case we also found that pcR was capable of outperforming clR, that pcR was preferred over NCN and that pcR and CMS performance showed no significant difference. The only qualitative difference between the results for the connected digit recognition task and those reported here is the fact that clR performed significantly better than NCN in the case of connected digit recognition. However, this may be explained by the small number of different phoneme contexts in the Dutch digit vocabulary: based on 18 monophones the number of different phone contexts is as low as 32.

One may expect that the deteriorating effect of introducing the left context dependency by using clR is a function of the number of different contexts in the vocabulary: The larger the number of different contexts, the larger this effect. In the case of the connected digits the number of different contexts is small. As a result the balance is still positive between the performance gain due to the channel bias removal on the one hand and the performance loss due to enhancement of the left context dependencies while using CI-HMMs on the other: For connected digit recognition clR outperforms NCN. In addition, there is a gain by applying the phase correction but it is small.

We determined the number of different phone contexts observed in the training set for the continous speech recognition task. Based on 37 monophones we found 2373 different phone contexts. When compared to the digit recognition task this is more than 70 times larger. As a consequence, the loss in recognition performance due to enhancement of the left context dependencies will be more important. Apparently, in our continuous speech recognition task this effect is now so large that it has become more important than the gain due to the channel bias removal. Therefore, for medium and large vocabulary continuous speech recognition clR does not improve recognition performance compared to NCN, while the gain is substantial when switching from clR to pcR or to CMS. Summing up, we obtain consistent results for two independent recognition tasks that differ considerably.

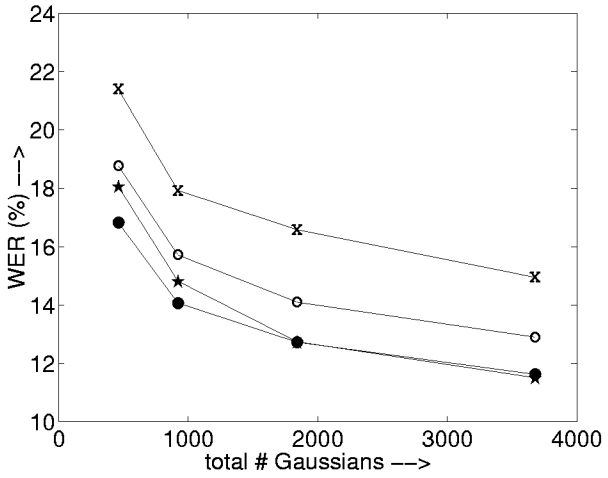

Figure 2: Recognition accuracy for clR (×), pcR (⋆) and CMS(●), compared to the feature set without CN (○) when using CI-HMMs.

5.2. Results for CD-HMMs

We repeated the comparison of CN techniques using CD-HMMs instead of CI-HMMs. The average WER results as a function of the total number of Gaussian densities are shown in Figure 3 for CD-HMMs with up to 8 Gaussians per state.

When using clR in the case of CD-HMMs one would expect that the loss of recognition performance due to enhancement of the left context is diminished, because different contexts are modelled with different states. When every individual left context could be modelled independently this effect would be at its maximum strength and the net result would be that one only has the gain in recognition performance due to the channel bias removal. However, in addition to the CD phone units there are a number of phone segments representing left contexts that are clustered during the state-tying.

In our CD-HMMs we used 167 CD units for modeling different left contexts. This is 14.2 times smaller than the total number of different phone contexts present in our training data, but 4.4 times larger than the number of different units we used to model the left contexts in our CI-HMM models. Based on these figures, one would expect that the performance of clR for our CD-HMMs would be more effective than in the case of CI-HMMs. As can be seen in Figure 3 the difference between clR and NCN has become smaller than the one we observed for CI-HMMs. For our best CI-HMMs we found that clR decreases recognition performance by 16% when compared to NCN. In the case of our best CD-HMMs the performance only drops 9%. Thus we indeed observe some gain when switching from CI- to CD-HMMs in the case of clR, but this improvement is limited due to the state-tying mechanism that we used to avoid undertraining. In fact the gain is too small such that we still do not benifit from the channel bias removal of clR in this case. This suggests that increasing the number of CD-HMM units could maybe further reduce the negative effect of combining different left contexts to the extent that clR would eventually outperform NCN.

As a second result it can be seen in Figure 3 that introducing the phase-correction immediately brings the recognition performance curve very close to the one for CMS (except at the models corresponding to 1 Gaussian per state). For the CD-HMMs corresponding to 8 Gaussians per state WER is improved by 15% when clR is replaced by pcR. This is in good agreement with the results of pcR obtained for CI-HMMs.

The present set of experiments indicates that a successful CN method should not introduce any phase distortion if CD-HMMs are used and the training data is not sufficient to model the left context dependency for all relevant contexts. This result is in good agreement with the conclusions in [4, 5, 6].

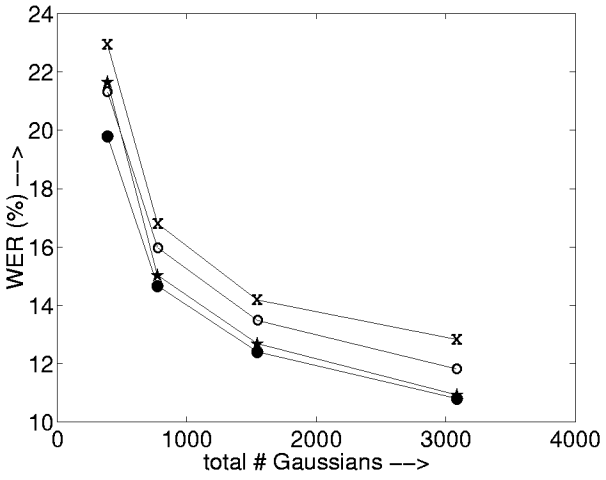

Figure 3: Recognition accuracy for clR (×), pcR (⋆) and CMS(●), compared to the feature set without CN (○) when using CD-HMMs.

6. CONCLUSIONS

We compared the performance of clR, CMS, pcR and using no channel normalization within the context of a medium vocabulary continuous speech recognition task. CMS over the whole utterance consistently gave the best results, but the difference with pcR is not sigificant. No CN yields better results than clR, due to the left context dependency introduced by the classical RASTA filter. Context dependent HMMs appear to reduce the detrimental effect of the left context dependency to some extent, but not enough to bridge the gap that separates it from CMS and pcR completely. Apparently, modeling artefacts of the RASTA filter is not the most effective use of limited amounts of training data. Finally, the conclusions of this study are in good agreement with the ones obtained in the context of connected digit recognition [4, 5, 6]. This underlines the importance of the phase response of CN filters, in addition to their magnitude response.

ACKNOWLEDGEMENT

This research was carried out within the framework of the Priority Programme Language and Speech Technology (TST). The TST-Programme is sponsored by NWO (Dutch Organisation for Scientific Research).

7. REFERENCES

1. S. Furui, 'Cepstral analysis technique for automatic speaker verification', IEEE Trans. Acoust. Speech Signal Process., ASSP-29, pp. 254-272, 1981.
2. H. Hermansky & N. Morgan, 'RASTA processing of speech', IEEE Trans. Speech Audio, 2(4), pp. 578-589, 1994.
3. J-C. Junqua, D. Fohr, J-F. Mari, T.H. Applebaum & B.A. Hanson, 'Time derivatives, cepstral normalization and spectral parameter filtering for continuously spelled names over the telephone' in Proc. Eurospeech-95, pp. 1385-1388, 1995.
4. J. de Veth & L. Boves, 'Channel normalization using phase-corrected RASTA', in Proc. ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, pp. 119-122, 1997.
5. J. de Veth & L. Boves, 'Phase-corrected RASTA for automatic speech recognition over the phone', in Proc. ICASSP-97, pp. 1239-1242, 1997.
6. J. de Veth & L. Boves, 'Channel normalization techniques for automatic speech recognition over the telephone', accepted for publication in Speech Communication, 1998.
7. V. Steinbiss, H. Ney, X. Aubert, S. Besling, C. Dugast, U. Essen, D. Geller, R. Haeb-Umbach, R. Kneser, H.-G. Meier, M. Oerder and B.-H. Tran, 'The Philips Research system for continuous-speech recognition', Philips J. Res., 49, pp. 317-352, 1995.
8. H. Strik, A. Russel, H. van den Heuvel, C. Cucchiarini and L. Boves, 'Localizing an automatic inquiry system for public transport information', in Proc. ICSLP-96, pp. 853-856, 1996.
9. S. Young & P. Woodland, 'HTK v1.4 User Manual', Speech Group, Cambridge University Engineering Department, UK, 1992.
10. H. Hermansky, 'Should recognizers have ears?', in Proc. ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, pp. 1-10, 1997.