COMPARISON OF CHANNEL NORMALISATION TECHNIQUES FOR AUTOMATIC SPEECH RECOGNITION OVER THE PHONE

Johan de Veth (1) & Louis Boves (1,2)

(1) Department of Language and Speech, University of Nijmegen,
P.O. Box 9103, 6500 HD Nijmegen, THE NETHERLANDS
(2) KPN Research, P.O. Box 421, 2260 AK Leidschendam, THE NETHERLANDS

ABSTRACT

We compared three different channel normalisation (CN) methods in the context of a connected digit recognition task over the phone: ceptrum mean substraction (CMS), RASTA filtering and the Gaussian dynamic cepstrum reprsentation (GDCR). Using a small set of context-independent (CI) continuous Gaussian mixture hidden Markov models (HMMs) we found that CMS and RASTA outperformed the GDCR technique. We show that the main cause for the superiority of CMS compared to RASTA is the phase distortion introduced by the RASTA filter. Recognition results for a phase-corrected RASTA technique are identical to those of CMS. Our results indicate that an ideal cepstrum based CN method should (1) effectively remove the DC-component, (2) at least preserve modulation frequencies in the range 2-16 Hz and (3) introduce no phase distortion in case CI HMMs are used for recognition.

1. INTRODUCTION

For automatic speech recognition over telephone lines it is well-known that recognition performance can be seriously degraded due to the transfer characteristics of the communication channel. In order to reduce the influence of the linear filtering effect of the telephone handset and telephone line, different channel normalisation (CN) techniques have been proposed [for example 1,2,3,4]. Several studies addressed the question of the relative effectiveness of different CN approaches [for example 5,6]. These studies were often limited to the extend that it was only established *which* CN technique was to be preferred. In this paper, we focus on the question *why* one approach is preferred over another.

We studied three different CN techniques in the context of a connected digit recognition task: cepstrum mean substraction (CMS) [1], RASTA filtering [2,3], and the Gaussian dynamic cepstrum representation (GDCR) [4]. For this task, we used hidden Markov models (HMMs) with Gaussian mixture densities describing the output probability density function of each state. Because we focussed attention on the question of what makes a CN technique a succesful one, we did not investigate the use of different types of acoustic parameter representations. Rather, we resticted ourselves to mel-frequency cepstral coefficients, log energy and their first time-derivatives.

This paper is further organised as follows. In section 2 we describe our feature extraction method. Next, in section 3, the telephone database that we used for our experiments is discussed. The topology of the HMMs, the way we performed training with cross-validation and the recognition syntax during testing are described in section 4. The recognition experiments are discussed in section 5. We will focus on the phase distortion introduced by the RASTA technique as this is the key difference between RASTA and CMS. We will show that removal of the phase distortion of the RASTA filter leads to a significant increase of recognition performance when using CI HMMs. Finally, in section 6 we sum up the main conclusions.

2. SIGNAL PROCESSING

Speech signals were digitized at 8 kHz and stored in A-law format. After conversion to a linear scale, preemphasis with factor 0.98 was applied. A 25 ms Hamming analysis window that was shifted with 10 ms steps was used to calculate 24 filterband energy values for each frame. The 24 triangular shaped filters were uniformly distributed on a mel-frequency scale. Finally, 12 mel-frequency cepstral coefficients (MFCC's) were derived. We did not apply liftering, because we were using continuous Gaussian mixture density HMMs with diagonal covariance matrices [7]. In addition to the twelve MFCC's we also used their 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. Feature extraction was done using HTK v1.4 [8]. We applied three CN techniques to the twelve MFCC coordinates of the feature vector in this paper. We either used RASTA with integration factor 0.98 [2,3], or the GDCR approach [4] or CMS [1]. We kept the original values of delta-MFCC's, logE and delta-logE.

3. DATABASE

The speech material for this experiment was taken from the Dutch POLYPHONE corpus [9]. Speakers were recorded over the public switched telephone network in the Netherlands. Handset and channel characteristics are not known; especially handset characteristics are known to vary widely. Among other things, the speakers were asked to read a connected digit string containing six digits. We divided this set of digit strings in two parts. For training we reserved a set of 960 strings, i.e. 80 speakers (40 females and 40 males) from each of the 12 provinces in the Netherlands (denoted trn960

Table 1: Phonemic transcriptions (column 2) and the number of realisations (columns 3 till 7) of each digit.

digit	transcription	trn960	trn480	tst911	tst240
nul	n Y l	590	294	548	136
een	e n	590	286	562	165
twee	t w e	591	296	597	181
drie	d r i	597	299	574	155
vier	v i r	569	284	523	135
vijf	v Ei f	573	273	526	124
zes	z E s	578	301	536	136
zeven	z e v Q n	582	270	510	130
acht	a x t	554	297	525	151
negen	n e x Q n	534	281	556	121

in short). An independent set of 911 utterances (tst911; 461 females, 450 males) was set apart for testing. (In principle we again wanted to have 40 female and 40 male speakers from each of the 12 provinces, but the very sparsely populated province of Flevoland provided only 21 female and 10 male test speakers). For proper initialisation of the models, we manually corrected automatically generated begin- and endpoints of each utterance in the trn960 data set. We did not always use all training and testing material. Most of the time, we used only half the amount of training data (i.e. 480 utterances, trn480; 240 females, 240 males). For cross-validation during training we used a subset of 240 utterances taken from the test set (tst240; 120 females, 120 males). For evaluation of the models when training was completed we always used the full test set tst911. We list the number of available realisations of each digit for all of our data sets in columns 3 till 6 of Table 1.

4. MODELS

4.1. Model topology

The digit set of the Dutch language was described using 18 context independent (CI) phone models (see second column of Table 1). Furthermore, we used four models to describe silence, very soft background noise, other background noise and out-of-vocabulary speech, respectively. Each CI model consists of a three state, left-to-right HMM, where only self-loops and transitions to the next state are allowed. The emission probability density functions are 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 containing 1, 2, 4, 8, 16 and 32 Gaussians for the emission probability density function of each state.

4.2. Training and recognition

The CI phone models were initialised starting from a linear segmentation within the boundaries taken from the hand-validated word segmentations. After this initialisation, an embedded Baum-Welch re-estimation was used to further train the models. Starting with a single Gaussian emission probability density function for each state, 20 Baum-Welch iterations were conducted; the models resulting from each iteration cycle are stored. Next, the optimal number of iterations was determined using the tst240 data set. For the set of models with the best recognition rate, the number of Gaussians was doubled and again 20 embedded Baum-Welch re-estimation iterations were performed. This process of training with cross-validation was repeated until models with 32 Gaussians per state were obtained.

During cross-validation as well as during recognition with data set tst911, the recognition syntax allowed for zero or more occurrences of either silence or very soft background noise or other background noise or out-of-vocabulary speech in between each pair of digits. At the beginning and at the end of the digit string one or more occurrences of either silence or very soft background noise of other background noise or out-of-vocabulary speech were allowed.

5. EXPERIMENTS

5.1. Comparison three CN methods

We trained models with up to 32 Gaussians per state using data set trn480. Four different sets of feature vectors were used to assess the effectiveness of CN: no CN, RASTA CN, GDCR CN and CMS CN. The best performing model sets according to the cross-validation data set tst240, were evaluated using test set tst911. The proportion of digits correct (i.e. the number of digits correctly recognized divided by the total number of digits in the test set) is shown as a function of the number of Gaussians per state in Figure 1. For the amount of test digits that we used, the 95% confidence interval is $0.6\%, 0.5\%, 0.4\%$ at a proportion of digits correctly recognized of $0.96, 0.97, 0.98$ respectively.

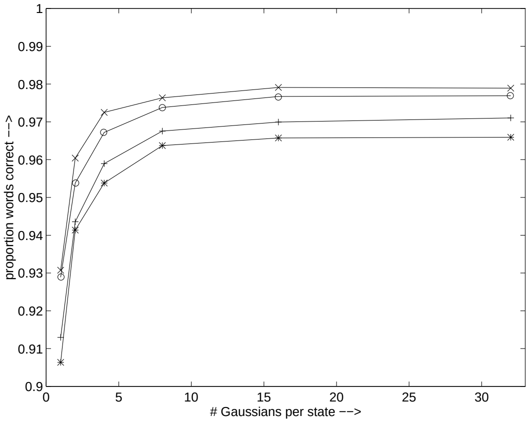

Figure 1: Recognition performance for four CN approaches: $\times =$ CMS, $O =$ RASTA, $+ =$ GDCR, $* =$ no CN.

Figure 1 clearly indicates that CN improves the recognition performance for each acoustic resolution that we tested. The improvements relative to the system without CN are significant at the 95% confidence level in case of RASTA and CMS, but they are not for GDCR. Notice further that the recognition performance increases monotonically as a function of the acoustic resolution in all four cases. Note, however, that in all cases the improvements are not significant for 16 and 32 Gaussians per state. In other words, 8 Gaussians per state appears to be sufficient for our connected digit recognition task. As a consequence, two different regions may be discerned on the acoustic resolution scale according to Figure 1. In the region up to 8 Gaussians per state recognition performance may be increased by either increasing acoustic resolution or applying a CN technique like RASTA or CMS. Above 8 Gaussians per state, however, increasing acoustic resolution does not result in any significant performance increase, whereas CN is still effective.

Using the RASTA filtered acoustic feature vectors, we conducted an experiment to verify that we used enough training data. To this aim models were trained with the trn960 data set. We did not observe a significant change in recognition performance. Therefore, we concluded that data set trn480 was indeed large enough.

5.2. RASTA vs. CMS in the time domain

According to the results in Figure 1, it appears that the different CN techniques that we studied can be ordered as follows: CMS $>$ RASTA $>$ GDCR $>$ no CN, where we used the symbol ”$>$” to indicate better CN effectiveness. The question is now of course: How can we understand this ordering? In [7], we argued that the RASTA filter frequency response preserves modulation frequencies in the maximally sensitive region of human auditory perception (2-16 Hz, [10]) much better compared to GDCR, especially in the region below 5 Hz. This preservation of modulation frequencies may very well explain the superiority of CMS and RASTA over GDCR. In order to see what causes the difference in recognition performance between RASTA and CMS, (which is significant at the 95 % level for systems with 2 and 4 Gaussians per state), we will take a detailed look at the effects of both techniques in the time domain.

We consider the signal shown in the upper panel of Figure 2 (we took a synthetic signal instead of a real MFCC coordinate time series for didactic purposes). The signal is a sequence of seven stationary segments (”speech states”) preceded and followed by a rest state (”silence”). Notice that the signal contains a constant overall DC-component (representing the effect of the communication channel). The RASTA filtered version of this signal is shown in the middle panel of Figure 2. Two important observations can be made. First, the DC-component has been effectively removed (at least for times larger than, say, 70 frames). Second, the shape of the signal has been altered.

With regards to the shape distortion we remark the following. First, the seven speech states of the signal that had a constant amplitude are now no longer stationary. Instead, the amplitude for each state shows a tendency to drift towards zero. Thus: RASTA filtering steadily decreases the value of cepstral coefficients in stationary parts of the speech signal, while the values immediately after an abrupt change are preserved. This explains the observation that the dynamic parts in the spectrogram of a speech signal are enhanced by RASTA filtering the cepstral coefficients [3,6]. As a consequence of this drift, however, a description of the signal in terms of stationary states with well-located means and small variances becomes less accurate. Second, the mean amplitude of each state has become a function of the state itself as well as the amplitudes of states immediately preceding it. This is the well-known left-context dependency introduced by the RASTA filter [3,11]. Because the absolute ordering of signal amplitudes is lost, states can no longer be straightforward identified by their mean amplitude (compare speech states two, four and seven before and after RASTA filtering in the upper and middle panel of Figure 2). For this reason, RASTA is less well suited when using CI models [cf. the remarks in 11]. Finally, we mention a third aspect of the shape distortion for completeness (which we feel is much less important though). Due to the small attenuation of high-frequency components, abrupt amplitude changes are smoothed.

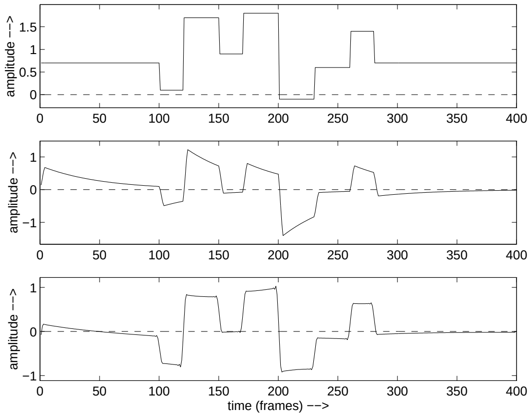

Figure 2: Synthetic signal representing one of the cepstral coefficients in the feature vector. Upper panel: Original signal containing a time-invariant DC-offset. Middle panel: RASTA filtered signal. Lower panel: Phase corrected RASTA filtered signal.

CMS has only one effect in the time domain: the DC-component is removed while the signal shape is exactly preserved (the signal is simply shifted as a whole). So, maybe the significant difference in performance between CMS and RASTA might be explained by the preservation of shape in the time domain in case CMS was used.

5.3. Phase correction for RASTA

In order to test this, we conducted a recognition experiment with an extended version of the RASTA filtering technique. We used the method decribed in [12] to do a phase correction on each MFCC coefficient after the RASTA filter was applied. We choose the phase correction such that the frequency dependent phase shift of the RASTA filter was exactly compensated, while at the same time preserving the original magnitude response of the RASTA filter by using an all-pass filter. The effect of the phase correction is shown in the lowest panel of Figure 2. As can be seen, the shape of the phase-corrected RASTA filtered signal resembles the shape of the original signal much better compared to the RASTA filtered signal. The phase correction (1) removes the amplitude drift towards zero in stationary parts of the signal and (2) removes the left-context dependency. In other words, phase-corrected RASTA (1) does not feature enhanced spectral dynamics and (2) is probably better suited for CI modeling.

We replaced the twelve MFCC’s by twelve phase-corrected RASTA filtered MFCC’s and trained new models using the same data sets trn480 and tst240 for training and cross-validation as before. Finally, we established the recognition performance using test set tst911. The results are shown in Figure 3, together with our previous results for CMS and RASTA. Figure 3 clearly shows that the performance of the phase-corrected RASTA features is identical to the CMS performance. Therefore, we conclude that our hypothesis was correct: The most important difference between CMS and RASTA is the phase distortion introduced by the RASTA filter, which is reflected in the time domain as a shape distortion of the signal. If the RASTA filter is adapted such that its phase distortion is exactly compensated while at the same time preserving the original magnitude response, the recognition performance becomes identical to the performance for CMS.

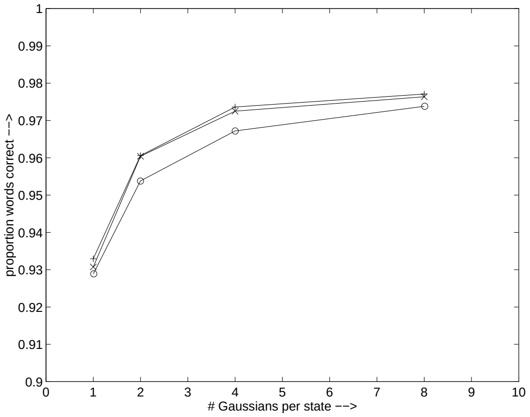

Figure 3: Recognition results for CMS ($\times$), RASTA (O) and phase-corrected RASTA ($+$).

We also conclude the following. It has been often suggested [3,6] that RASTA techniques provide better recognition performance *because* the spectral dynamics are enhanced. Our analysis shows that this enhancement is caused by the phase distortion of the RASTA filter. When we removed the phase distortion, we removed the enhancement of spectral dynamics. However, the recognition performance did not go down in our experiments (on the contrary). Therefore, the argument should be reconsidered that the success of RASTA filtering techniques should be attributed to the enhancement of spectral dynamics. Our experiments suggest that removal of the DC-component is the most important feature of RASTA.

Finally, taking our findings for CMS, RASTA and GDCR together, we can formulate three constraints that an ideal cepstrum based CN technique should satisfy: (1) the DC-component should be effectively removed, (2) the magnitude response should be preserved in the range of 2-16 Hz, which is the maximally sensitive region of human auditory perception, and (3) the technique should not introduce any phase distortion when combined with CI modeling.

6. CONCLUSIONS

We compared three different CN methods in the context of a connected digit recognition task over the phone. Using a small set of CI continuous Gaussian mixture HMMs, we showed that CMS and RASTA outperform the GDCR technique. Furthermore, we showed that the main cause for the superiority of CMS compared to RASTA is the phase distortion introduced by the RASTA filter. The recognition results for a phase-corrected RASTA technique were identical to those of CMS. Our results suggest that the ability of RASTA to effectively remove the DC-component is more important than the enhancement of spectral dynamics.

Acknowledgement

This work was funded by the Netherlands Organisation for Scientific Research (NWO) as part of the NWO Priority programme Language and Speech Technology.

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, A. Bayya & P. Cohn , ’Compensation for the effect of the communication channel in auditory-like analysis of speech’, in Proc. Eurospeech-91, Genova, Sept. 1991.

3 H. Hermansky & N. Morgan , ’RASTA processing of speech’, IEEE Trans. Speech Audio, 2(4), pp. 578-589, 1994.

4 K. Aikawa, H. Singer, H. Kawahara & Y. Tohkura , ’A dynamic cepstrum incorporating time-frequency masking and its application to continuous speech recognition’, in Proc. ICASSP-93, pp. 668-671, 1993.

5 J-C. Junqua, D. Fohr, J-F. Mari, T.H. Applebaum & B.A. Hanson , ’Time derivatives, cepstral normalisation and spectral parameter filtering for continuously spelled names over the telephone’, in Proc. Eurospeech-95, pp. 1385-1388, 1995.

6 H. Singer, K.K. Paliwal, T. Beppu & Y. Sagisaka , ’Effect of RASTA-type processing for speech recognition with speaking-rate mismatches’, in Proc. Eurospeech-95, pp. 487-490, 1995.

7 P. Boda, J. de Veth & L. Boves , ’Channel normalisation by using RASTA filtering and the dynamic cepstrum for automatic speech recognition over the phone’, to appear in Proc. ESCA Workshop on the Auditory Basis of Speech Perception, Keele, July 1996.

8 S. Young & P. Woodland , ’HTK v1.4 User Manual’, Speech Group, Cambridge University Engineering Department, UK, 1992.

9 E.A. den Os, T.I. Boogaart, L. Boves & E. Klabbers , ’The Dutch Polyphone corpus’, in Proc. Eurospeech-95, pp. 825-828, 1995.

10 R. Drullman, J.M. Festen & R. Plomp , ’Effect of temporal envelope smearing on speech reception’, J. Acoust. Soc. Am., vol. 95, pp. 1053-1064, 1994.

11 J. Cohen , Final report of the chairman, Frontiers of Speech Processing - Robust Speech Recognition 93, 1993.

12 M. J. Hunt , ’Automatic correction of low-frequency phase-distortion in analogue magnetic recordings’, Acoustic Letters, vol. 32, pp. 6-10, 1978.