PHASE-CORRECTED RASTA FOR AUTOMATIC SPEECH RECOGNITION OVER THE PHONE

Johan de Veth *Louis Boves*

A2RT, Department of Language and Speech, University of Nijmegen, Nijmegen, THE NETHERLANDS

ABSTRACT

In this paper we propose an extension to the classical RASTA technique. The new method consists of classical RASTA filtering followed by a phase correction operation. In this manner, the influence of the communication channel is as effectively removed as with classical RASTA. However, our proposal does not introduce a left-context dependency like classical RASTA. Therefore the new method is better suited for automatic speech recognition based on context-independent modeling with Gaussian mixture hidden Markov models. We tested this in the context of connected digit recognition over the phone. In case we used context-dependent hidden Markov models (i.e. word models), we found that classical RASTA and phase-corrected RASTA performed equally well. For context-independent phone-based models, we found that phase-corrected RASTA can outperform classical RASTA depending on the acoustic resolution of the models.

1. INTRODUCTION

For automatic speech recognition (ASR) over the telephone it is well-known that the recognition performance may be seriously degraded due to the transfer characteristics of the handset microphone and the telephone channel [1]. In order to reduce the influence of the linear filtering effect of the communication channel, different channel normalisation (CN) techniques have been proposed (for example [2, 3, 4]). In our paper we present a new, extended version of the classical RASTA filtering technique [3].

Classical RASTA filtering features two important properties: (1) attenuation at low modulation frequencies and (2) enhancement of the dynamic parts of the spectrogram [3]. The first property explains why classical RASTA filtering is such an effective method for CN: In the cepstral or log-energy domain, linear filtering by a quasi-stationary communication channel gives rise to an additive constant bias term [1]. The attenuation at low modulation frequencies effectively removes this DC-component. It has been suggested that the second property is also benificial for good recognition performance [3]. Recently, it was shown that the enhancement of the dynamic parts of the spectrogram obtained by classical RASTA represents a crude approximation of the effects of temporal forward masking in human auditory perception [5, 6]. Thus, classical RASTA may be viewed as a combination of CN and a crude model of human auditory time-masking.

The method we propose consists of classical RASTA filtering followed by a phase correction operation. The phase correction is chosen such that the frequency-dependent non-linear phase-shift of the classical RASTA filter is compensated, while at the same time preserving the original magnitude response of the classical RASTA filter [7]. In this manner phase-corrected RASTA effectively removes the influence of the communication channel and at the same time does not enhance the dynamic parts of the spectrogram (i.e. does not model human auditory time-masking). In addition, phase-corrected RASTA removes the well-known left-context dependency introduced by classical RASTA. Therefore, one may expect that the new CN method is better suited for ASR based on context-independent (CI) modeling.

This paper is organised as follows. In section 2 we describe details of the phase-corrected RASTA method. We will focus on the non-linear phase distortion introduced by classical RASTA and describe the method we used to restore the original phase. Next, in section 3, the signal processing for our experiments is described. The telephone database that we used for our experiments is discussed in section 4. After this, the topology of the hidden Markov models (HMMs), the way we performed training with cross-validation and the recognition syntax during testing are described in section 5. The results of our recognition experiments are discussed in section 6. As we will see, these experiments 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 7 we sum up the main conclusions.

2. PHASE-CORRECTED RASTA

Consider the signal shown in the upper panel of Figure 1 (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 1. 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.

In *Proc. ICASSP-97, Apr 21-24, Munich, Germany*, pp. 1239-1242, 1997

With regards to the shape distortion the following can be noticed. 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[3]. 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]. Because the absolute ordering of signal amplitudes is lost, states can no longer be straightforwardly characterised by their mean amplitude (compare speech states two, four and seven before and after RASTA filtering in the upper and middle panel of Figure 1). For this reason, RASTA is less well suited when using CI models (cf. the remarks in [3]). Finally, we mention a third aspect of the shape distortion for completeness (which we feel is less important though). Due to the small attenuation of high-frequency components, abrupt amplitude changes are smoothed.

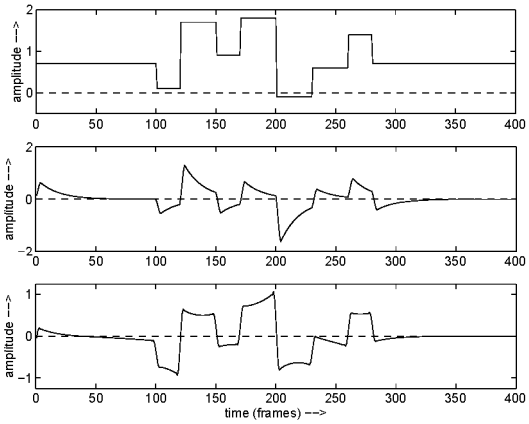

Figure 1: Synthetic signal representing one of the cepstral coefficients in the feature vector. Upper panel: Original signal containing a time-invariant DC-offset. Middle panel: RASTA filtered signal. Lower panel: Phase corrected RASTA filtered signal.

The complex frequency response of the classical RASTA filter $H_R(\omega)$ may be written as

$$H_R(\omega) = |H_R(\omega)|\, e^{j\phi_R(\omega)}, \tag{1}$$

with ω the modulation frequency (in radians), $|H_R(\omega)|$ the RASTA magnitude response and $\phi_R(\omega)$ the RASTA phase response. The log-magnitude and phase response of the classical RASTA filter with integration factor $a = -0.94$ are shown in Figures 2a,b for modulation frequencies in the range $0 - 20$ Hz. This range includes the $2 - 16$ Hz region, which has been shown to be most important for good recognition by humans [8]. From Figure 2b, it can be seen that the phase response is non-linear for modulation frequencies below approximately 3 Hz. As we will see, the non-linear phase response of the classical RASTA filter is the main cause of the shape distortions observed in the middle panel of Figure 1.

In order to compensate the phase distortion of the RASTA filter, while at the same time preserving the original magnitude response, we followed the procedure suggested in [9]. After the classical RASTA filter, an all-pass filter can be applied such that its phase response $\phi_{pc}(\omega)$ is exactly the opposite of the phase response of the RASTA filter

$$\phi_{pc}(\omega) = -\phi_R(\omega). \tag{2}$$

Thus, we obtain for the frequency response $H_{pc}(\omega)$ of the phase-correction filter

$$H_{pc}(\omega) = e^{-j\phi_R(\omega)}. \tag{3}$$

Applying this phase correction after the classical RASTA filter, we have for the frequency response $H_{pcR}(\omega)$ of the complete phase-corrected RASTA filter

$$H_{pcR}(\omega) = H_R \times H_{pc} = |H_R(\omega)|\,. \tag{4}$$

We implemented the phase correction filter $H_{pc}(\omega)$ in practice as a pole-zero filter. Thus, we solved for coefficients {b,a} that satisfy

$$e^{-j\phi_R(\omega)} = \frac{b_0 + b_1 e^{-j\omega} + \ldots + b_q e^{-jq\omega}}{1 + a_1 e^{-j\omega} + \ldots + a_p e^{-jp\omega}}, \tag{5}$$

where q (p) is the order of the numerator (denominator) polynomial. We used a standard fitting procedure of Matlab with $q = 1$ and $p = 7$ to calculate the {b,a} coefficients [10]. Because the resulting pole-zero filter is unstable, we applied the inverse of this filter to the time-reversed signal after which a second time-reversal operation was performed.

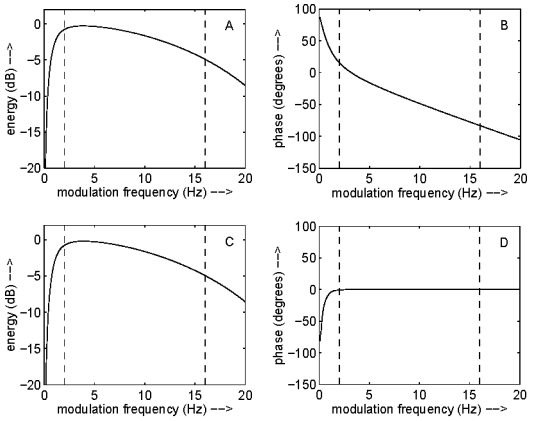

Figure 2: A. Log-magnitude response classical RASTA. B. Phase response classical RASTA. C. Log-magnitude response phase-corrected RASTA. D. Phase response phase-corrected RASTA.

Figures 2c,d show the log-magnitude response and the phase response of the phase-corrected RASTA filter. It can

be seen that the new log-magnitude response is equal to the original one and at the same time that the new phase curve is flat in the region of important modulation frequencies.

The result of applying the phase correction filter in the time-domain is shown in the lowest panel of Figure 1. 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. In order to test the second point, we compared classical and phase-corrected RASTA using context-dependent (CD) and context-independent HMMs.

3. 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. 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 [11].

Because we wanted to focus on the difference between phase-corrected and classical RASTA, we did not investigate the use of other types of acoustic parameter representations. We applied the CN techniques to the twelve MFCC coordinates of the feature vector in this paper. We used RASTA with integration factor -0.94 [3] and the corresponding phase-corrected RASTA method. We kept the original values of delta-MFCC's, logE and delta-logE.

4. DATABASE

The speech material for this experiment was taken from the Dutch POLYPHONE corpus [12]. 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. None of the utterances used for training or test had a high background noise level.

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 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. For most of the CI models 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.

5. MODELS

5.1. Model topology

The digit set of the Dutch language was described using either 18 CI phone models or 10 word-based (i.e. CD) models. In addition, we used four models to describe silence, very soft background noise, other background noise and out-of-vocabulary speech, respectively. Each CI model consisted of three states. Each CD model contained exactly the same number of states as were used for the word in the CI description. In this manner, the number of states for a CD digit model ranged between 9 and 15. The total number of different states describing the digit HMMs was 99 for the CD models and 56 for the CI models. All HMMs were left-to-right, 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 and 16 Gaussians for the emission probability density function of each state.

5.2. Training and recognition

The models were initialised starting from a linear segmentation within the boundaries taken from the hand-validated 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 were 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 16 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 or other background noise or out-of-vocabulary speech were allowed.

6. EXPERIMENTS

In a first set of two experiments we trained CI HMMs for classical RASTA with integration factor -0.94 (in short: clR(-0.94)) and the corresponding phase-corrected RASTA (pcR(-0.94)) using train set trn480. We used test set tst911

to determine the recognition accuracy of both CN methods as a function of the acoustic resolution. The accuracy was defined as the one minus the quotient of the sum of the number of substitutions, insertions and deletions, and the total number of digits. The results are shown in Figure 3. It can be seen that pcR(-0.94) performed significantly better than clR(-0.94) when 8 and 16 Gaussians per state were used. For 1, 2 and 4 Gaussians per state both methods were equivalent. In [7] we reported a similar comparison using integration factor -0.98 and found that pcR(-0.98) was significantly better for 2, 4 and 8 Gaussians per state. At the time of this writing, an explanation for this difference between our current and previous experiments remains an open issue. Of these four different CN methods, best results overall were obtained for pcR(-0.94) with 16 Gaussians per state.

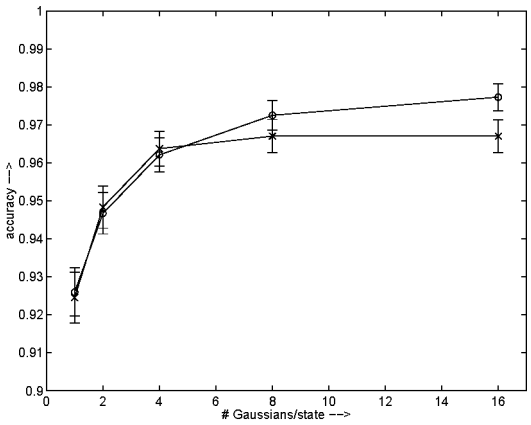

Figure 3: Recognition accuracy for RASTA (X) and phase-corrected RASTA (O) using CI HMMs.

In a second set of experiments, we trained CD HMMs for clR(-0.94) and pcR(-0.94). We observed effects of undertraining when we used train set trn480 for these models. Therefore, we doubled the amount of training data and used train set trn960. For the whole range of acoustic resolutions studied, we found that both methods performed equally well within the 95% confidence regions. This is what one would expect, because performance of CD HMMs will not suffer from the left context dependency introduced by clR.

Finally we note the following. It has been suggested [3] that classical RASTA provides better recognition performance because DC-components are effectively removed and because the spectral dynamics are enhanced. Our analysis shows that the enhancement of spectral dynamics is caused by the phase distortion of the RASTA filter. When we removed the phase distortion, we removed the enhancement of spectral dynamics. However, we did not observe a degradation of recognition performance in our CI experiments. Our experiments suggest that removal of the DC-component offered by classical RASTA is more important than enhancement of spectral dynamics.

7. CONCLUSIONS

We have proposed a new extension to the classical RASTA CN technique. In our proposal the classical RASTA filter is followed by an all-pass phase correction filter. In this manner the left-context dependency introduced by the classical RASTA filter is removed, while at the same time DC-components are still as effectively removed. Experiments using CI HMMs for connected digit string recognition over the phone, suggest that phase-corrected RASTA can outperform classical RASTA, depending on the combination of the integration factor and the number of Gaussians per state. Best results so far were obtained with the combination of phase-corrected RASTA(-0.94) and 16 Gaussians per state. In addition, 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] H. Hermansky, N. Morgan, A. Bayya & P. Kohn, 'Compensation for the effect of the communication channel in auditory-like analysis of speech', in Proc. Eurospeech-91, 1991.

[2] S. Furui, 'Cepstral analysis technique for automatic speaker verification', IEEE Trans. Acoust. Speech Signal Process., ASSP-29, pp. 254-272, 1981.

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

[4] 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.

[5] H. Hermansky & M. Pavel, 'Psychophysics of speech engineering systems', in Proc. ICPhS-95, pp. 3.42-3.49, 1995.

[6] H. Hermansky, 'Auditory modeling in automatic recognition of speech', ESCA Workshop on the Auditory basis of speech perception, Keele University (UK), 15-19 July, 1996.

[7] J. de Veth & L. Boves, 'Comparison of channel normalisation techniques for automatic speech recognition over the phone', in Proc. ICSLP-96, pp. 2332-2335, 1996.

[8] 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.

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

[10] J. H. Little & L. Shure, 'Matlab Signal Processing Toolbox Users Guide', The MathWorks, Inc., May 1993.

[11] S. Young & P. Woodland, 'HTK v1.4 User Manual', Speech Group, Cambridge University Engineering Department, UK, 1992.

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