A comparison of LPC and FFT-based acoustic features for noise robust ASR

Febe de Wet, Bert Cranen, Johan de Veth, Loe Boves

Department of Language and Speech
University of Nijmegen, The Netherlands
{F.de.Wet, B.Cranen, J.deVeth, L.Boves}@let.kun.nl

Abstract

Within the context of robust acoustic features for automatic speech recognition (ASR), we evaluated mel-frequency cepstral coefficients (MFCCs) derived from two spectral representation techniques, i.e. the fast Fourier transform (FFT) and linear predictive coding (LPC). ASR systems based on the two feature types were tested on a digit recognition task using continuous density hidden Markov phone models. System performance was determined in clean acoustic conditions as well as in different simulations of adverse acoustic conditions. The LPC-based MFCCs outperformed their FFT counterparts in most of the adverse acoustic conditions that were investigated in this study. A tentative explanation for this difference in recognition performance is given.

1. Introduction

ASR systems that operate in 'real world' conditions must function in various different acoustic conditions. However, many of these 'real world' acoustic conditions do not resemble the acoustics represented by the training data. If training data are clean (in the sense that care is taken to avoid excessively noisy recording environments), noise that is present at recognition time will result in a mismatch between training and test conditions. Such mismatches lead to a degradation in recognition performance.

The eventual goal of our research is to reduce the negative impact of training-test mismatch on recognition performance. Using acoustic features that are inherently less sensitive to noise, is one of the ways in which this may be accomplished. In a previous investigation on robust acoustic features for ASR, FFT-based MFCCs were compared with acoustic features derived from an LPC estimate of speech spectra [2]. In that study, it was found that LPC-based acoustic features are inherently more robust to at least some kinds of background noise than their FFT counterparts. In the current investigation, we elaborate on this comparative study in an attempt to determine which properties of the LPC-based features constitute their robustness.

To this aim, the recognition performance of LPC and FFT-based ASR systems were measured in clean as well as adverse acoustic conditions. Adverse acoustic conditions were simulated by adding car, babble and factory noise from the Noisex CD [6] to clean speech data at different signal-to-noise ratios (SNRs). It should be kept in mind that the observations made in simulated noise conditions may not always generalise to 'real world' applications - clean signals with artificially added noise are by no means exact representations of 'real world' noise conditions. For instance, they do not capture the way in which people tend to change their rate and manner of speaking in noisy acoustic conditions (Lombard effect [5]). Nevertheless, these simulations are widely used for experimental purposes, e.g. [4], because they provide a framework within which recognition performance in clean and noisy acoustic conditions may easily be compared. Such a framework also provides the possibility to measure the impact of additive noise on the statistical properties of the data at acoustic feature level.

In the experiments reported on in this paper, the degree of mismatch between the training and test data was varied from being perfectly matched to completely mismatched. *Perfectly matched* refers to an experimental set-up where the acoustic conditions during training and recognition were identical. *Completely mismatched*, on the other hand, refers to experiments where the acoustic models were trained on clean data while recognition was performed on noisy data. Recognition performance was also determined for two intermediate conditions. In the first of these, the non-speech models were trained on noisy data that matched the noise at recognition time while the speech sound models were trained on clean data and tested in noise (*speech mismatch*). In the second intermediate experiment, all the acoustic models were trained on noisy data that matched the noise at recognition time, except for the non-speech models. These were trained on clean data and subsequently used to perform recognition in noisy conditions (*non-speech mismatch*).

2. Experimental set-up

2.1. Clean speech material

The speech material for our experiments was taken from the Dutch POLYPHONE corpus [3]. Speech was recorded over the public switched telephone network in the Netherlands, using a primary rate ISDN interface and a sampling frequency of 8 kHz. The POLYPHONE corpus contains various examples of (read) speech utterances. Only the connected digit items were used in our current investigation. The number of digits in each string varied between 3 and 16. A set of 1,997 strings (16,582 digits) was used for training. Care was taken to balance the training material with respect to gender, region (an equal number of speakers from each of the 12 provinces in the Netherlands) and the number of tokens per digit. 504 digit string utterances (4,300 digits) were used for cross-validation during training. An independent test set of 1,008 utterances (8,300 digits) was used for evaluation. The cross-validation and independent test sets were balanced according to the same criteria as the training material.

2.2. 'Noisified' speech material

Recognition performance was evaluated under three different simulations of adverse acoustic conditions. Car, babble and factory noise from the Noisex CD were chosen as the noise conditions for the current experiments. For all practical purposes, the car and babble noise signals may be considered as stationary. The factory noise signal contains a number of ham-

mer blows and could therefore be considered as an example of non-stationary noise. In terms of their long time average spectra, both babble and factory noise can be classified as (almost) broad-band noise. In contrast, the car noise is band limited to very low frequencies (below 250 Hz).

The Noisex signals contain broad-band frequency information while the information content of the signals in our database is limited to the frequency range of the public switched telephone network in the Netherlands. As an approximation of the channel's frequency response, the Noisex signals were band-pass filtered before they were added to the clean signals[1]. The addition was performed such that the SNR level of the resulting signals was 10 dBA.

2.3. Acoustic pre-processing

Figure 1 gives a graphical overview of the acoustic pre-processing that was implemented in our experiments. A pre-emphasis factor of 0.98 and a 25ms Hamming window shifted with 10ms steps were used to prepare the data for spectral analysis. Two spectral representations of each data frame were subsequently obtained: The first was derived by calculating a 256 point FFT. The second was based on an LPC spectral estimate of the spectral envelope of an AR filter resulting from a 10^{th} order LPC analysis. In order to provide a representation of the speech signal which is as similar as possible for the two spectral estimation techniques, we reconstructed the spectral envelope from the 10 LPC coefficients and used the residual signal energy to scale them back to their original energy level.

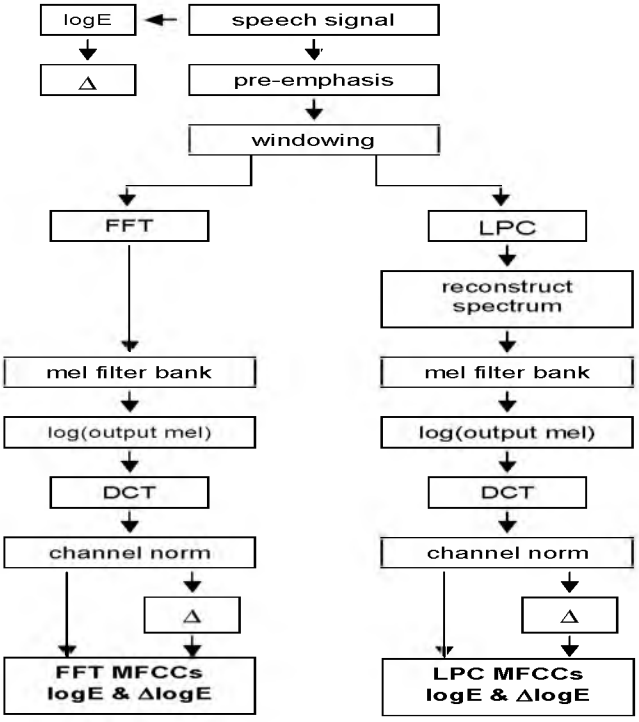

Figure 1: *The acoustic pre-processing procedure.*

From both the LPC and the FFT-based spectra 16 mel-scaled log-energy values were calculated. The filters in the mel bank were triangularly shaped, half overlapping and uniformly distributed on a mel-frequency scale between 0 and 2143.6 mel, corresponding to 0-4000 Hz on a linear frequency scale. 12 MFCCs were derived from the log of the mel bank outputs using the Discrete Cosine Transform. Cepstral mean subtraction (CMS) was applied as a channel normalisation technique. We used the off-line version of the CMS algorithm, i.e. the cepstral mean was calculated per utterance. The first derivatives of the MFCCs were also computed and added to the vector of 12 channel normalised feature values. The HTK normalised log-energy and delta log-energy values of each frame were also included in the acoustic feature vectors.

2.4. Hidden Markov Modelling

Continuous density hidden Markov models (HMMs) were used to describe the statistics of the speech sounds. The ten Dutch digit words were described in terms of 18 *phone* models. Two additional models were used to represent the statistical properties of the silence and background noise (non-speech) in the recordings of the POLYPHONE database. Each phone unit was represented as a left-to-right HMM of three states. Only self-loops and transitions to the next state were allowed. All HMMs were implemented using diagonal covariance matrices and 16 Gaussian mixtures components per state. HTK2.1 was used for training and testing.

Four sets of acoustic models were trained: One set on clean training data and three sets on the training data noisified with car, babble and factory noise, respectively. The acoustic models were all trained using cross-validation. The recognition syntax used during cross-validation and testing allowed for digit strings varying in length from 3 to 16 digits to be recognised, without prior knowledge of the length of a particular string. The syntax also allowed silence and noise to be recognised between consecutive digits as well as at the beginning and the end of each utterance.

3. Results

All recognition results are given in terms of Word Error Rate (WER) defined as:

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

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.

3.1. Perfectly matched acoustic conditions

Table 1 gives an overview of the WERs that were measured when the acoustic conditions during training and recognition were identical. The 95% confidence intervals of the WERs are given in brackets.

Acoustic condition	FFT	LPC
clean	3.7 (0.4)	3.8 (0.4)
car noise	6.0 (0.5)	6.3 (0.5)
babble noise	10.5 (0.7)	12.5 (0.7)
factory noise	12.9 (0.7)	12.9 (0.7)

Table 1: *WERs (with 95% confidence intervals) for the LPC and FFT-based ASR systems in perfectly matched conditions.*

These values show that, with the exception of the WERs measured in babble noise, there is no significant difference be-

[1]Passing the car noise signal through the 'channel' filter effectively reduced it to broad-band noise, because the high energy peak in the low part of the spectrum was removed, and the remaining energy was almost uniformly distributed over the frequencies in the pass-band.

tween the recognition rates achieved by the LPC and FFT-based ASR systems. Compared to the recognition rate in clean acoustic conditions, both systems suffer a degradation in recognition performance in the presence of noise. Car noise seems to represent the mildest of the three noise conditions while the largest performance loss is observed in factory noise.

3.2. Completely mismatched acoustic conditions

The WERs obtained in completely mismatched acoustic conditions (all models trained on clean data and tested in noise) are summarised in Table 2. The values in the 'times' (×) column of the table are the factor by which the WERs increased in going from the perfectly matched condition to the completely mismatched condition.

noise	FFT			LPC		
	match	mis	×	match	mis	×
car	6.0	26.6	4.4	6.3	13.1	2.1
babble	10.5	35.1	3.3	12.5	24.9	2.0
factory	12.9	41.0	3.2	12.9	25.4	2.0

Table 2: *WERs for the LPC and FFT-based ASR systems in perfectly matched (match) and completely mismatched (mis) acoustic conditions.*

The results in Table 2 show that the HMMs trained on LPC MFCCs in the clean condition yield WERs that are significantly lower than those produced by the corresponding FFT-based system in car, babble and factory noise. Similar experiments were conducted at SNRs of 5 and 15 dBA. The LPC-based system outperformed its FFT-based counterpart in both instances.

In order to determine whether this difference in recognition performance could best be explained in terms of the recognition of speech sounds or non-speech sounds, two more experiments were conducted: One in which only the speech models were mismatched and one in which only the non-speech models were mismatched.

3.3. Speech mismatch

The aim of this experiment was to determine to what extent the differences in recognition performance observed in Table 2 may be explained by *mismatched speech sound models*. The speech sound models were therefore *trained* on clean data and subsequently *tested* in noise. The non-speech models, on the other hand, were *trained and tested* on noisy data. Table 3 gives an overview of the WERs that were measured in this experiment. The values in the 'times' (×) column have the same meaning as in Table 2.

noise	FFT			LPC		
	match	sp mis	×	match	sp mis	×
car	6.0	17.9	3.0	6.3	11.3	1.8
babble	10.5	27.8	2.6	12.5	22.6	1.8
factory	12.9	29.4	2.3	12.9	23.3	1.8

Table 3: *WERs for the LPC and FFT-based ASR systems in perfectly matched (matched) and speech mismatched (sp mis) acoustic conditions.*

According to the results in Table 3, the FFT-based ASR system suffers a substantial loss in recognition performance on account of the mismatched speech models. The LPC-based ASR system also performs worse than in the perfectly matched condition, but not to the same extent as the FFT-based system.

3.4. Non-speech mismatch

In this experiment, the speech sound models were *trained and tested* on noisy data while the non-speech models were *trained* on clean data and subsequently *tested* in noisy conditions. The aim of the experiment was to determine the impact of *mismatched non-speech models* on the recognition rates of the two systems. The resulting WERs are summarised in Table 4. The values in the 'times' (×) column have the same meaning as before.

noise	FFT			LPC		
	match	Nsp mis	×	match	Nsp mis	×
car	6.0	8.1	1.4	6.3	9.5	1.5
babble	10.5	15.0	1.4	12.5	17.2	1.4
factory	12.9	15.2	1.2	12.9	14.5	1.1

Table 4: *WERs for the LPC and FFT-based ASR systems in perfectly matched (match) and non-speech mismatched (Nsp mis) acoustic conditions.*

The WERs in Table 4 reveal that the performance of both systems degrades on account of the mismatched non-speech models. The two systems seem to be equally sensitive to the mismatch - the factor by which the WERs increase is almost the same for both systems in all three noise conditions.

4. Discussion

In Table 1, it was observed that the LPC and the FFT-based ASR systems perform equally well in almost all perfectly matched acoustic conditions, but worse in noise than in clean conditions. This result is not un-expected since, for both representations it holds that, if the training data is 'clean', the statistical properties of the acoustic feature vectors are mainly determined by the characteristics of speech, and the resulting HMMs represent the statistical properties of speech sounds only. However, if the training data is 'noisified', all signals have a common component, i.e. the noise. As a consequence, the modelling capacity of the HMMs trained on the noisy data will partially be spent on describing this common component. Their discriminative ability can therefore be expected to deteriorate, with a corresponding degradation in recognition performance.

The results presented in Section 3 showed that, in some instances, LPC-based MFCCs appeared to be more robust against mismatched training-test conditions than FFT-based MFCCs (cf. Tables 2 and 3). However, in some cases hardly any difference could be observed between the recognition rates of the LPC and FFT-based ASR systems (cf. Table 4). Moreover, for at least one type of noise the ASR system based on LPC MFCCs performed worse than its FFT counterpart (cf. Table 1). But where do these differences in recognition performance come from?

When noise is added to a signal, its distortional effect (in the log magnitude domain) is most evident in the spectral valleys while the spectral peaks remain relatively unchanged. The most prominent difference between LPC and FFT spectral estimators is related to the way in which they describe spectral peaks and valleys: The LPC estimation of a spectrum yields a spectral envelope with a relatively accurate description of the peaks in the spectrum. The representation of the valleys is restricted to a description of the energy level, no detailed information about their fine spectral structure is included. Except for a decrease in dynamic range, additive noise should therefore have little effect on LPC spectral estimates - at least as long as the spectral

properties of the noise do not include strong spectral peaks and speech-related spectral peaks are not completely masked by the noise level. In contrast, non-parametric descriptions of spectra, such as the FFT, describe spectral peaks and valleys in equal detail. If the spectral valleys are filled by additive noise, the spectral fluctuations introduced by the noise are described in just as much detail as the spectral peaks. As a consequence, FFT-based acoustic features, e.g. the cepstral coefficients that were used in this study, can be expected to suffer more from noise related variation than features derived from a parametric representation such as LPC.

According to the WERs in Table 1, the LPC-based ASR system is only outperformed by its FFT counterpart in the presence of babble noise. The spectral properties of babble noise are most probably the reason for this observation: Babble noise is often classified as broad-band noise based on its long time average spectrum. However, within an analysis window of 25ms, we have observed that it often exhibits spectral structure that closely resembles the spectral structure of speech sounds such as vowels. The LPC spectral estimate of a substantial number of data frames may have suffered on account of the spectral structure introduced by the babble noise. Since an LPC spectral estimator focuses its modelling power on the most prominent peaks in the spectrum, it is more susceptible to the impact of noise signals with strong spectral peaks than FFT spectral estimates are. The impact of noise-related peaks on LPC spectra will be especially detrimental if their energy level is comparable to the amount of energy in the speech-related peaks.

In the experiment where the HMMs trained on clean data were tested in the presence of noise, the recognition rate of the HMMs based on LPC MFCCs was superior to the performance of those trained on FFT MFCCs (see Table 2). Moreover, the WERs in Table 3 show that, if the speech sound models are mismatched, the corresponding drop in recognition rate is much larger for the FFT-based system than for the LPC-based system - even if the non-speech models are well-matched. This observation seems to suggest that the superior performance of the LPC-based system observed in Table 2 may primarily be ascribed to the speech sound models. We hypothesise that HMMs trained on 'clean' LPC-based MFCCs retain their ability to discriminate between speech sounds in the presence of noise, because the LPC MFCC vectors obtained in clean and noisy conditions are more similar than FFT MFCCs. The higher degree of similarity could be explained as follows: At 10 dBA the valleys in the clean spectra are filled up by noise, but not to the extent that the speech related peaks in the spectra are no longer distinguishable. Because LPC-based spectra focus on the spectral peaks and ignore details in the spectral valleys, the corresponding MFCCs are more robust to the introduction of non-speech related fluctuations in the spectral valleys than FFT-based MFCCs.

The results in Table 4 revealed that the LPC and FFT-based ASR systems are equally sensitive to the impact of mismatched non-speech models on recognition performance. The substantial increase in WER that is observed for both systems may be attributed to the fact that more than 40% of the data in our corpus is non-speech. The non-speech parts of the clean data are generally characterised by low energy levels. The non-speech frames will therefore be the first to have their spectral structure masked by the added noise. Under these circumstances the LPC-based MFCCs loose their advantage over their FFT counterparts - hence the equal performance of the two ASR systems in this experiment.

In order to determine to what extent our current observations may be generalised to other experimental conditions, we conducted similar experiments using the connected digit material in SpeechDat Car Italian and the aurora2.0 databases [1]. For both these databases the trends in the results were similar to those reported in Section 3, i.e. the two systems performed equally well in well-matched conditions, while the system based on LPC MFCCs outperformed the system based on FFT MFCCs in almost all mismatched conditions.

The only instance in which the FFT-based system proved to be superior, was for aurora2.0 data with SNRs of 5 dB and lower. This result may be related to the inferior performance of the LPC-based system that was observed in some of our own experiments (cf. Tables 1 and 4). Our explanation of these results was that LPC looses its advantage as a spectral estimator as soon as the spectral properties of the background noise obscure the underlying spectral properties of the speech in the data. This will most probably also be the case at very low SNRs because the speech-related peaks in the data are often completely masked by the level of the noise. As a consequence, an LPC estimator will no longer be able to make an accurate parametric fit of the data.

5. Conclusions

The results presented in this paper show that, at SNRs of 5dBA and higher, the LPC and FFT-based ASR systems perform equally well if the acoustic properties of training and test data are well-matched. The two systems also seem to be equally sensitive to a mismatch in non-speech data if the speech models are well-matched. However, if the speech sound models are mismatched, the system based on LPC MFCCs outperforms its FFT counterpart, even if the non-speech models are well-matched. It may therefore be concluded that acoustic features, MFCCs in this instance, based on an LPC spectral estimate are inherently more noise robust than FFT-based acoustic features. The difference in recognition performance that was observed in the various mismatch conditions could be explained by assuming that the major advantage of LPC-based features is their ability to ignore spectral details in spectral valleys. Future research will be aimed at quantifying the differences between the statistical properties of LPC and FFT-based acoustic features.

6. References

[1] L. Boves, D. Jouvet, J. Sienel, R. de Mori, F. Bechet, L. Fissore, and P. Laface. ASR for automatic directory assistance: the SMADA project. In *Proceedings of ASR 2000*, pages 249–254, Paris, France, 2000.

[2] F. de Wet, B. Cranen, J. de Veth, and L. Boves. Comparing acoustic features for robust ASR in fixed and cellular network applications. In *Proceedings of ICASSP 2000*, pages 1415–1418, Istanbul, Turkey, 2000.

[3] E. A. den Os, T. I. Boogaart, L. Boves, and E. Klabbers. The Dutch Polyphone corpus. In *Proceedings of Eurospeech '95*, pages 825–828, Madrid, 1995.

[4] H. G. Hirsch and D. Pearce. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proceedings of ASR 2000*, pages 181–188, Paris, France, 2000.

[5] J. C. Junqua. The influence of acoustics on speech production: A noise-induced stress phenomenon known as the lombard reflex. *Speech Communication*, 20:13–22, 1996.

[6] Noisex. NOISE-ROM-0. NATO: AC243/(Panel 3)/RSG-10, ESPRIT: Project 2589-SAM, 1990.