# COMPARING ACOUSTIC FEATURES FOR ROBUST ASR IN FIXED AND CELLULAR NETWORK APPLICATIONS

*Febe de Wet, Bert Cranen, Johan de Veth and Louis Boves*

$A^2RT$, Department of Language & Speech, University of Nijmegen
Nijmegen, The Netherlands
{F.de.Wet, B.Cranen, J.deVeth, L.Boves}@let.kun.nl

## ABSTRACT

Within the context of automatic speech recognition (ASR) applications for telephony, we investigate the acoustic preprocessing issues that are at stake in going from the fixed line to the cellular network. Because the spectral representation used in enhanced full rate GSM is linear prediction, we investigate the relative advantages and drawbacks of conventional mel-frequency cepstral coefficient (MFCC) parameters derived from a non-parametric fast Fourier transform (FFT) and MFCC parameters derived from a linear predictive coding (LPC) spectral estimate. Robust Formant parameters, also derived from an LPC description of the spectrum, are studied as an alternative to MFCCs. Within the framework of connected digit recognition based on hidden Markov models, ASR performance was measured for clean conditions, as well as for three different additive noise conditions. In addition, the performance of a conventional recognition procedure was compared with the performance of an ASR system based on our *Acoustic Backing-off* implementation of Missing Feature Theory (MFT).

## 1. INTRODUCTION

Over the last years much effort has been spent to improve the robustness of ASR against adverse acoustic conditions. Most of the work has been done in the context of telephone systems. Although it has long been known that acoustic background noise levels in cellular calls are generally higher than in calls on the fixed network, few studies have explicitly addressed the issues that set cellular networks apart from fixed networks. An important difference between fixed and cellular is the signal processing in the cellular networks. GSM coding implies an LPC analysis at the source, combined with parameter quantization. In the Enhanced Full Rate (EFR) coder the LPC parameters are coded in the form of Line Spectral Frequencies (LSFs) [6].
The current paper reports on the first steps towards taking our previous work in robust ASR from the fixed to the cellular environment. Our approach to increasing noise robustness in ASR is closely related to MFT, implemented in the form of Robust Statistical Pattern Recognition. In previous papers the implementation was named *Acoustic Backing-off*, to indicate the mathematical basis for dealing with outliers in the feature value distributions due to distortions in the speech signal [2]. MFT provides a framework for analyzing and interpreting the properties of different acoustic feature representations with respect to their robustness to adverse conditions. In prior reports we have provided arguments in support of the hypothesis that feature representations which guarantee that distortions which are local in the time-frequency space remain local, should have an advantage over representations which disperse local distortions over the entire feature vector [4, 3]. In our previous investigations on the properties of feature representations all features were derived from spectral estimates obtained with an FFT analysis. FFT-based spectral estimates exhibit the "locality" property that should keep local distortions local. However, EFR GSM coders entail spectral estimates obtained by means of LPC analysis. LPC spectral estimates are based on a parametric model of the speech signal. Since LPC spectral estimates are limited to a fixed number of parameters in an all-pole model, local distortions which cause narrow peaks or steep slope changes due to extra noise energy in a part of the spectrum may affect the estimate of the envelope in remote frequency regions. Thus, otherwise similar feature representations derived from LPC spectral estimates might exhibit other "locality" properties than their counterparts based on FFT analysis.

The EFR GSM codec represents the spectral envelope in the form of LSFs. One reason to select LSFs for coding is their "locality" property: if one LSF is damaged (for instance, due to radio transmission problems), the effects of the distortion will not propagate to other frequency regions. The same reasoning suggests to use LSF parameters in ASR on speech from an EFR GSM source. Unfortunately, despite early promising results [7], it appeared that LSFs have other properties which make them less suitable for ASR. For instance, the spectral interpretation of individual LSF parameters depends strongly on the presence or absence of close neighbors. Therefore, the interest in LSFs for application in ASR has subsided [8].

The fact that the EFR GSM codec is based on LSFs inspired us to investigate potential transformations of LSF parameters that would be suitable for ASR. One such transform that has been successfully applied in speech synthesis, is known as *Robust Formants* (RF) [9]. Since the RF representation is directly related to LSFs, it should retain the desirable "locality" property which is important for robustness against radio transmission errors. Furthermore, RFs have a consistent spectral interpretation, which means that they might overcome the problems that plagued previous attempts to use LSFs in ASR. On the other hand, the po-

tential advantage of RFs may be annihilated if the underlying LPC spectral estimate yields parameters that are all biased because of a narrow band of additive noise.

In this paper we investigate the relative advantages and drawbacks of MFCC parameters derived from a non-parametric FFT, MFCC parameters derived from an LPC-based spectral estimate and RF parameters. This three-way comparison is carried out under clean conditions, as well as under three different additive noise conditions. These comparisons should allow us to clarify the inherent properties of the different parameter representations. As an additional factor in the research we compare the performance of a conventional recognition procedure with the performance of a decoder based on our *Acoustic Backing-off* implementation of Missing Feature Theory [2]. The latter comparison should shed light on the degree to which different noises and different feature representations combine to yield distorted features for which the marginalization approach to MFT shows an advantage.

## 2. EXPERIMENTAL SET-UP

### 2.1. Speech Material

The speech material for the experiments was taken from the Dutch POLYPHONE corpus [5]. Speech was recorded over the public switched telephone network in the Netherlands, using a primary rate ISDN interface. Among other things, the speakers were asked to read several connected digit strings. The number of digits in each string varied between 3 and 16. A set of 1997 strings (16582 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 (4300 digits) were used for cross-validation during training. For evaluation an independent test set of 1008 utterances (8300 digits) was used. The cross-validation and independent test sets were balanced according to the same criteria as the training material. None of the utterances used for training or cross-validation testing had a high background noise level.

### 2.2. Simulating Adverse Acoustic Conditions

Recognition performance was evaluated under three different simulations of adverse acoustic conditions. In each case noise signals were added to the speech such that the resulting SNR was 10 dBA[1]. The first type of noise that we studied was band limited white noise. The band limited signal was obtained by filtering a white noise signal with a fifth order elliptical filter. The cut-off frequencies of the band-pass filter ($F_{low}$ = 833 Hz, $F_{high}$ = 1446 Hz) were chosen such that approximately one quarter of the mel-frequency range would be affected by the noise. Babble and factory noise were chosen as examples of broad band noise. These noise signals were obtained from the Noisex CD.

[1]Both the speech and noise energy levels were weighted according to the A-scale.

### 2.3. Acoustic Features

In order to make a fair comparison between the FFT and the LPC descriptions of the spectrum, all acoustic features were derived directly from speech signals and not from EFR encoded speech. A 25 ms Hamming window shifted with 10 ms steps and a pre-emphasis factor of 0.98 was applied. After windowing and pre-emphasis, the data was processed as described in the following paragraphs.

To obtain the FFT-MFCCs, the data were converted to the frequency domain by applying the FFT. In the frequency domain 16 filter band log energy values were calculated. The filter bands were triangularly shaped, half overlapping and uniformly distributed on a mel-frequency scale (covering 0-2143.6 mel; this corresponds to the linear range of 0-4000 Hz). 12 mel cepstra were computed from the filter bank outputs using the Discrete Cosine Transform (DCT). Cepstral mean subtraction (CMS) was applied as a channel normalization (CN) technique. The time derivatives of the FFT-MFCCs were also computed and added to the vector of 12 channel normalized feature values. The log-energy and delta log-energy values of each frame were also included in the 26-dimensional acoustic feature vectors.

In determining the LPC-MFCCs, the frequency domain description of the signals was based on the spectral envelope of an AR filter resulting from a $10^{th}$ order LPC analysis (in correspondence with EFR GSM encoding standards [6]). The rest of the calculation procedure was exactly the same as for the FFT-MFCCs, i.e. log mel-frequency filter banks, DCT, etc. The LPC-MFCC system was therefore also based on 26-dimensional acoustic feature vectors.

The Robust Formant algorithm that we used was developed within the framework of speech synthesis. The algorithm always finds a specified number of spectral maxima, called "Formants" for historical reasons, in such a way that the resulting formant tracks are continuous from frame to frame [9]. The algorithm was used to extract 5 formant($F_i$), bandwidth($B_i$) pairs for each speech frame, also based on a $10^{th}$ order LPC analysis. During experimentation we did not use the bandwidth values directly, but transformed them to their corresponding Q-values, where the Q-values were determined as:

$$Q_i = \frac{F_i}{B_i} \tag{1}$$

The 5 {formant,Q-value} pairs were combined with their first time derivatives and their corresponding log-energy and delta log-energy values to obtain 22-dimensional acoustic feature vectors. No additional normalization was applied to the RF parameters.

### 2.4. Hidden Markov Modeling

Continuous density hidden Markov models (HMMs) were used to describe the statistics of the speech sounds. A phone-based system was used, i.e., the basic speech sounds that were to be recognized were phones. The ten Dutch digit words were described with 18 phone models. Three additional models were used to capture the statistical properties of the silence, background noise and out-of-vocabulary speech in the recordings of the POLYPHONE database. Each phone unit was represented as a left-to-right HMM consisting of three states. Only self-loops and transitions

to the next state were allowed. The total number of different states was 63 (54 for the phones plus 9 for the noise models); 16 Gaussians per state were trained. HTK2.1 was used for training and testing the HMMs [10]. Training was done according to the cross-validation scheme described in [1]. All HMMs were implemented using diagonal covariance matrices[2] and each model set was trained only once, using clean speech data, i.e., with undisturbed features. The recognition syntax used during cross-validation and testing allowed for digit strings varying in length from 3 to 16 digits to be recognized, without prior knowledge of the length of a particular string.

### 2.5. Recognition procedure

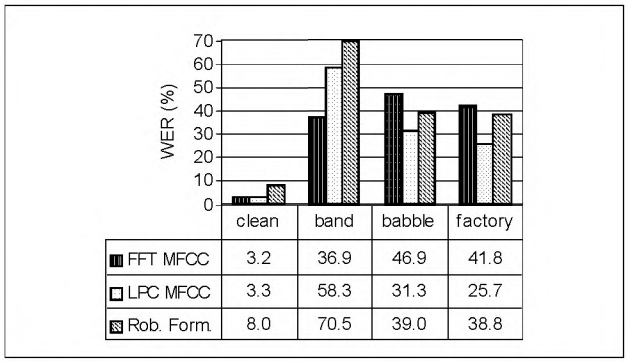
Conventionally, the contribution of each Gaussian distribution to the local distance function used during dynamic programming is calculated as a quadratic function over the entire feature value range. In [2] it was proposed to make the outer ends of this quadratic function constant. This was shown to be equivalent to modeling the feature value observations by means of two distributions: one obtained from the training data and another, uniform distribution which represents all feature values not seen during training. The weight assigned to either distribution can be varied so as to increase or decrease the contribution of the unseen values. This strategy was called *Acoustic Backing-off*. In the current investigation we performed recognition experiments both with and without the application of *Acoustic Backing-off*.

## 3. RESULTS

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

$$WER = \frac{S + D + I}{N} \times 100\%. \qquad (2)$$

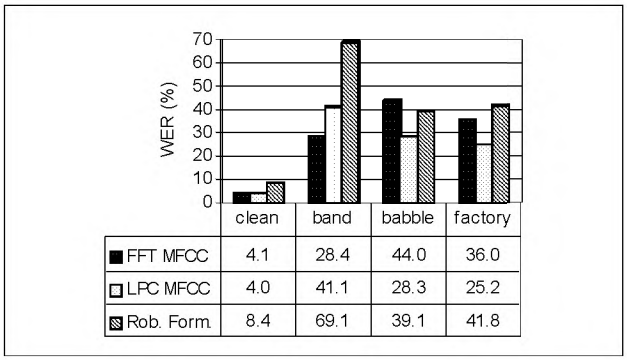
N is the total number of words in the test set, S denotes the total number of substitution errors, D the total number of deletion errors and I the total number of insertion errors.




| | clean | band | babble | factory |
|---|---|---|---|---|
| FFT MFCC | 3.2 | 36.9 | 46.9 | 41.8 |
| LPC MFCC | 3.3 | 58.3 | 31.3 | 25.7 |
| Rob. Form. | 8.0 | 70.5 | 39.0 | 38.8 |

Figure 1: WER results for a conventional recognition procedure.

[2]The correlation between the RF features was small enough to justify the use of a diagonal covariance matrix.




| | clean | band | babble | factory |
|---|---|---|---|---|
| FFT MFCC | 4.1 | 28.4 | 44.0 | 36.0 |
| LPC MFCC | 4.0 | 41.1 | 28.3 | 25.2 |
| Rob. Form. | 8.4 | 69.1 | 39.1 | 41.8 |

Figure 2: WER results for a recognition procedure with Acoustic Backing-off.

The baseline recognition performance for the three different feature sets in the clean condition is shown in column 1 of Figure 1. The WERs for the FFT-MFCC and LPC-MFCC systems are virtually the same, i.e. 3.2% and 3.3%. The RF system performs significantly worse, achieving a WER of 8.0%. However, it should be kept in mind that the system has fewer parameters than the two other systems (22 components as opposed to the 26 of the MFCC vectors) and no explicit channel normalization.

Figure 1 also gives an overview of the recognition performance of the three different ASR systems under adverse acoustic conditions using a conventional recognition procedure. Comparing the FFT-MFCC results with those obtained for the LPC-MFCCs shows that the FFT-based system achieves a lower WER in the band limited noise condition than the LPC-based parameters. On the other hand, the LPC-MFCCs outperform the FFT-MFCC system in the broad band noise conditions, i.e., babble and factory noise. The LPC-MFCC system performs consistently better than the RF system in all noise conditions. However, in the presence of broad band noise, the RF system appears to be more robust than the FFT-MFCC system.

Figure 2 summarizes the WERs achieved by the same ASR systems under the same adverse acoustic conditions using a recognition procedure with Acoustic Backing-off. The application of Acoustic Backing-off in noisy conditions leads to lower WERs in all cases, but two. These exception are the WERs obtained for the RF system in babble and factory noise. The corresponding loss of recognition performance in the clean condition was kept within the range of 1.0% absolute. Despite the general improvement in system performance, the observations made for the "conventional" recognizer still hold true: FFT-based features do better in band limited noise while LPC-based features yield better results in broad band noise conditions. The LPC-MFCC system also keeps its advantage over the RF system.

## 4. DISCUSSION & CONCLUSIONS

Since there is almost no difference between the performance of the FFT-MFCC and the LPC-MFCC systems in the clean condition, our results indicate that the ability of the FFT to properly represent spectral valleys is of very little

consequence for ASR in clean conditions. In broad-band noise conditions, however, the LPC-MFCCs yield lower WERs than their FFT counterparts. This result suggests that, in broad-band noise conditions, features based on smoothed spectra, LPC-MFCCs in this instance, are more robust than those based on spectral representations with a higher resolution such as FFT-MFCCs. This observation can be explained by the fact that LPC spectral estimates are mainly determined by the peaks in the spectrum. Therefore, filling the valleys with broad band noise should have only a small effect on the estimates of the envelope. Features derived from an FFT, on the other hand, should suffer from relatively large differences in the depth of the valleys.

The opposite holds true for band limited noise, where the FFT-MFCCs perform better than the LPC-MFCCs. This is probably due to the fact that the addition of band limited noise at 10 dBA has a substantial influence on the frequency positions where strong spectral slope changes are observed. LPC parameters are known to be extremely sensitive to such abrupt changes in the spectrum. Due to the fact that the spectral envelope must be described with a fixed number of parameters, all parameters will suffer from this type of distortion. The "misleading" spectral peak/plateau will therefore lead to an incorrect parametric description of the speech information. In FFT analysis, on the other hand, the spectral description of the signal information in the unaffected frequency regions will remain intact in the presence of band limited noise.

The different behavior of LPC and FFT derived features under narrow and broad band noise conditions holds irrespective of the application of *Acoustic Backing-off*. This result suggests that the choice between the two spectral estimators is not arbitrary. Unless one must anticipate narrow band distortions, LPC based spectral estimates may exhibit advantages that are supported by solid theoretical considerations. For the same theoretical reasons FFT based spectral estimates are more robust against narrow band noise.

Even though the acoustic features of both the LPC-MFCC and the RF systems are derived from a $10^{th}$ order LPC analysis, the LPC-MFCC system outperforms the RF system in all the experimental conditions that we investigated. The inferior performance of the RF system might be due to the fact that it has fewer parameters than the LPC-MFCC system. However, it is more likely that the RF parameters do not provide a representation of the spectral envelope that is suitable for HMM-based ASR. A partial explanation may be found in the intrinsic quantization of the bandwidths in the algorithm proposed in [9]. For use in ASR the fact that the algorithm always finds the same number of spectral maxima may also constitute a disadvantage: if the spectral envelope contains fewer clear peaks than the predefined number, it is difficult to predict the effect on the "superfluous" features. If anything, the bandwidths would tend to attain maximum values. This might result in distributions which are difficult to model with a Gaussian mixture. Given that RF features are not really suited to describe all the speech sounds in the system equally well, they do remarkably well in comparison with e.g. the FFT-MFCC system in broad band noise conditions. In future research, we will investigate whether the information contained in the MFCCs and the RF parameters is complementary and may therefore be combined to improve recognition performance.

Application of *Acoustic Backing-off* in the clean condition consistently leads to a slight loss of recognition performance. This is the price one has to pay for the difference between the distributions obtained during the training and the distributions used in the decoder [2]. Contrary to our previous findings, however, the gain from *Acoustic Backing-off* under adverse conditions is marginal at best; for the RF features it even causes a small degradation, compared to conventional decoding. For the time being we can only explain the failure of *Acoustic Backing-off* by assuming that the combinations of noise and feature representations used in this study results in a small proportion of outliers – the only type of distortion which can be handled effectively by MFT – and a much larger proportion of feature values that are distorted, but only to a degree that does not allow them to be discarded by our implementation of MFT.

## 5. REFERENCES


[1] J. de Veth and L. Boves. Channel normalization techniques for automatic speech recognition over the telephone. *Speech Communication*, 25:149–164, 1998.

[2] J. de Veth, B. Cranen and L. Boves. Acoustic backing off in the local distance computation for robust automatic speech recognition. In *Proceedings of ICSLP '98*, pages 1427–1430, Sydney, Australia, 1998.

[3] J. de Veth, B. Cranen, F. de Wet and L. Boves. Acoustic pre-processing for optimal effectivity of missing feature theory. In *Proceedings of Eurospeech '99*, pages 65–68, Budapest, Hungary, September 1999.

[4] J. de Veth, F. de Wet, B. Cranen and L. Boves. Missing feature theory in ASR: make sure you miss the right type of features. In *Workshop on robust methods for speech recognition in adverse conditions*, pages 231–234, Tampere, Finland, 1999.

[5] 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.

[6] ETSI. Draft ETSI EN 300 726 V6.0.0; Digital cellular systems (phase 2); Enhanced Full Rate (EFR) speech transcoding. http://www.etsi.org/, 1999.

[7] K. K. Paliwal. On the use of Line Spectral Frequency parameters for speech recognition. *Digital Signal Processing*, 2:80–87, 1992.

[8] T. F. Quatieri, R. B. Dunn, D. A. Reynolds and J. P. Campbell. Speaker and language recognition using speech coding parameters. In *Proceedings of Eurospeech '99*, pages 787–790, Budapest, Hungary, September 1999.

[9] L. F. Willems. Robust formant analysis. In *IPO Annual report 21*, pages 34–40, Eindhoven, The Netherlands, 1986.

[10] S. Young, J. Jansen, J. Odell, D. Ollason and P. Woodland. *The HTK Book (for HTK Version 2.1)*. Cambridge University, Cambridge, UK, 1995.