LSF ALIASING REMOVAL FOR IMPROVED SPEAKER VERIFICATION

Halil Uzuner, Stephane Villette, Khaldoon Al-Naimi, Ahmet Kondoz

Centre for Communication Systems Research (CCSR)
University of Surrey, Guildford, Surrey, GU2 7XH, UK
{h.uzuner, s.villette, k.al-naimi, a.kondoz}@surrey.ac.uk

ABSTRACT

Applications such as transaction authentication may require speaker recognition systems to operate on compressed speech transmitted over mobile phone networks. However, speech compression degrades speech quality, and hence causes a reduction in recognition performance. It has been shown that the classic technique for extraction of Line Spectral Frequency (LSF) parameters in speech coders is prone to aliasing distortion. The use of a low-pass filtering on upsampled LSF vectors has been shown to alleviate this problem, therefore improving speech quality. In this paper, the effect of this Non-Aliased LSF (NA-LSF) extraction method on speaker recognition performance is observed using GSM-EFR coded speech. When the NA-LSFs are used in the coder the verification performance loss introduced by the GSM-EFR is reduced, producing similar performance to uncoded speech.

1. INTRODUCTION

The increasing interest in mobile communications leads to a higher demand for speaker recognition applications which use coded speech. Typical application examples, where the coded speech is used in speaker recognition systems, include transaction authentication such as telephone banking, and law enforcement for identifying suspects. In these applications, there is a loss in speaker recognition performance due to the speech compression carried out by the speech coders. This recognition performance loss increases when the quality of the speech coder (i.e. bit rate) decreases [1, 2]. Different methods such as Score Normalisation [2] and Speech Coder Recognition [3] have been used to reduce the loss in speaker recognition performance using coded speech. In this paper, we introduce a method that operates within the speech coder without modifying its design. The experiments are performed using GSM-EFR [4] as it is the most widely used standard coder for mobile communications in Europe. Nevertheless, the results are also expected to be applicable to AMR [5], as its principles are similar to GSM-EFR. It has been shown in [6] that the LSF parameters contain high frequency variations, which cause some aliasing noise in the LSF parameters. These unwanted frequency components can be removed by employing an anti-aliasing filter. In this paper, the use of NA-LSF parameters in the GSM-EFR coder is shown to increase speaker recognition performance. The paper is organised as follows: Section 2 presents the problems associated with the classical LSF extraction methods and also introduces the anti-aliasing filtering process in the GSM-EFR coder to eliminate spectral aliasing. Section 3 describes the use of NA-LSF parameters in a speaker verification system and presents the experimental results. Concluding remarks are given in Section 4.

2. ANTI-ALIASING FILTERING PROCESS

2.1. LSF extraction from a decimation perspective

The GSM-EFR coder performs Linear Prediction (LP) analysis [7] twice for each speech frame using autocorrelation, utilising two asymmetric 30 ms wide analysis windows that are different in shape. These windows are designed in such a way that look-ahead delay is not required. The first window is constructed from the two halves of Hamming windows that have different sizes. The first window's weight is concentrated at the second subframe of the coder analysis window, bearing in mind that the GSM-EFR algorithm divides each analysis window into 4 subframes. The second window is constructed from a Hamming window and a quarter of a cosine function cycle. The weight of this window is concentrated at the fourth subframe. Each one of these windows is then used to produce 10 LSF vectors per speech frame. In order to analyse the effects of aliasing on LSF parameters, the LSF extraction is performed at a higher sampling rate (i.e. parameter extraction at every sample) than the system rate. LSF tracks show the LSF parameter evolution over time and they are obtained by plotting each parameter value in time using over-sampled LSF vectors. When down-sampling is performed on the LSF tracks at the system rate (i.e. the rate of vector transmission), the LSF vectors which are generated are identical to the original LSF extraction method. During the down-sampling process, any LSF track that contains spectral components at frequencies greater than half of its vector transmission frequency causes spectral overlapping. This produces some aliasing noise in the extracted LSF parameters. In

order to remove the high frequency variations observed in the LSF track spectra, a pre-processing stage was proposed in [6] that involves the use of low-pass filtering before LSF vector decimation. We employed this anti-aliasing filtering approach in the LSF parameter extraction section of the GSM-EFR. It will be shown that the use of NA-LSF parameters improve the quality of the synthesised speech produced by the GSM-EFR. Therefore, GSM-EFR speech coding with NA-LSF parameter extraction has been shown to provide more efficient speaker modeling and testing processes and ultimately better speaker recognition performance.

2.2. Low-pass filtering

In our experiments, the low-pass filtering has been applied as follows:

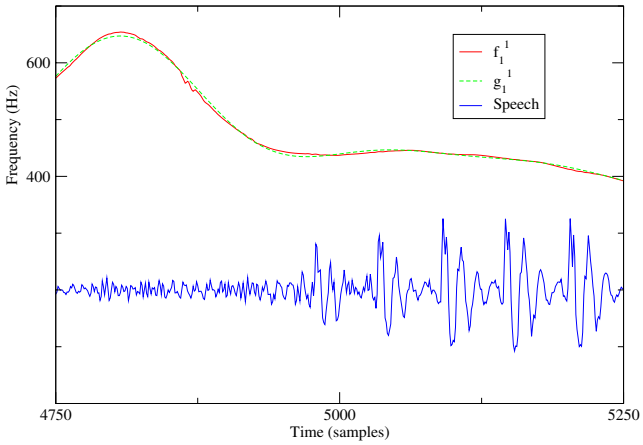
1. Extract two sets of LSF vectors ($\mathbf{f}^1(n) = f_1^1(n), \ldots, f_p^1(n)$ and $\mathbf{f}^2(n) = f_1^2(n), \ldots, f_p^2(n)$) from the two sets of LPC vectors ($\mathbf{l}^1(n) = l_1^1(n), \ldots, l_p^1(n)$ and $\mathbf{l}^2(n) = l_1^2(n), \ldots, l_p^2(n)$) computed at every sample for each analysis window of the GSM-EFR, where p is the LP filter order and n is time.

2. Construct two sets of LSF tracks using LSF vectors (f_p^1 and f_p^2) for each analysis window obtained from the first step.

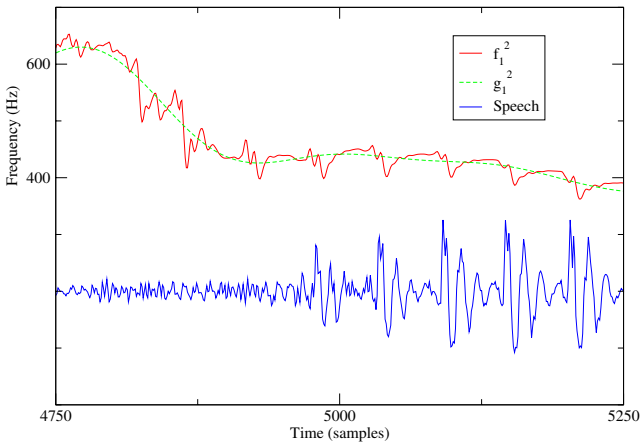
3. For each LSF track (f_p), perform low-pass filtering in the frequency domain with a cut-off frequency that is chosen according to the vector transmission rate.

In order to demonstrate the effect of low-pass filtering on the LSF tracks, sets of LSF tracks were obtained using speech samples collected from the 8 kHz down-sampled version of TIMIT database (TIMIT8k) [8].

Figure 1 (a),(b) and Figure 2 (a),(b) show the 1^{st} and 10^{th} LSF tracks of the original and the NA-LSF parameters obtained from the first and second LP analysis windows of the GSM-EFR, respectively. It can be observed from these figures that the NA-LSF tracks follow a smoother behaviour compared to the original tracks. The original LSF tracks contain a large amount of variation. This variation in the original LSF tracks are more prominent in intervals where there is a transition between voiced and unvoiced speech. The distortions on the LSF tracks are more evident with the higher order LSF parameters (i.e. 10^{th} LSF parameter). Also it can be observed in these figures that the amount of distortion on the original LSF tracks for the second LP analysis window is much higher compared to the the first LP analysis window. This is caused by the use of two different window compositions as described in Subsection 2.1. The LSF parameters of the second LP analysis window of the GSM-EFR are more distorted as a result of the weighting applied by this unusually-shaped asymmetric window.



(a) LSF tracks f_1 and g_1 for the first analysis window

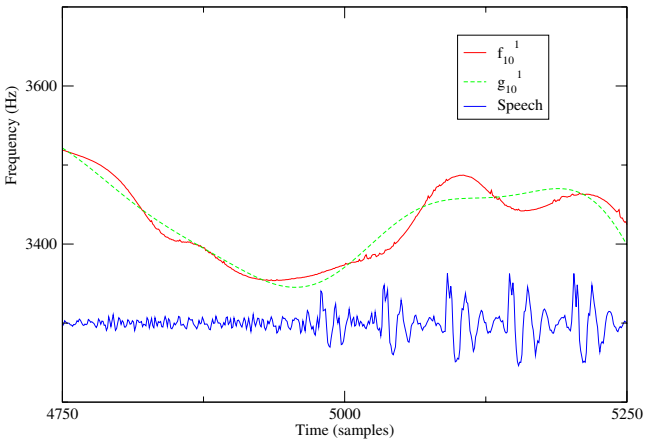


(b) LSF tracks f_1 and g_1 for the second analysis window

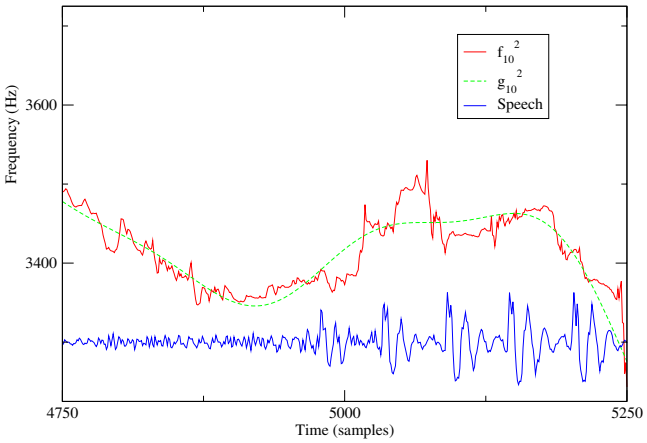
Fig. 1. Variations in the 1^{st} LSF track for original f_1 and low-pass filtered g_1 LSFs.

3. SPEAKER VERIFICATION USING NA-LSFs IN GSM-EFR

It is shown in [1, 2] that speaker verification performance degrades when coded speech is used for speaker training and testing processes. As a result of speech coding effects, the verification performance decreases under matched training and testing conditions (i.e. the training and the testing data are collected from the same coder) [2]. Under mismatched conditions (i.e. the training data are collected from the clean speech and the testing data are collected from the coded speech and vice versa), the verification performance degradation was found to be even higher due to the mismatch between the training



(a) LSF tracks f_{10} and g_{10} for the first analysis window



(b) LSF tracks f_{10} and g_{10} for the second analysis window

Fig. 2. Variations in the 10^{th} LSF track for original f_{10} and low-pass filtered g_{10} LSFs.

models and the test vectors [2, 3]. Different methods have been proposed in [1, 2, 3] to improve the recognition performance under matched and mismatched training and testing conditions. In this paper, both mismatched and matched conditions are used to demonstrate the benefit of the NA-LSF parameter extraction method. The following subsections describe the experimental setup of the NA-LSF extraction process for the GSM-EFR coder and the performance evaluation of the speaker verification system using the GSM-EFR coded TIMIT8k database (TGSM database) with NA-LSFs employed in the coder. The NA-LSF extraction is performed as described in Subsection 2.2. The FFT window size is chosen to be large enough in order to avoid the effects of the large side lobes of the rectangular window. The cut-off frequency of the low-pass filter used for the two LP windows is at 25 Hz as this corresponds to a 10 ms vector transmission rate of the GSM-EFR coder (keeping in mind that there are two LP analysis windows shifted by 20 ms every frame).

3.1. Speaker verification experiments

Speech databases NA-LSF TGSM and Org-LSF TGSM represent the GSM-EFR coded TIMIT database using the NA-LSF and original LSF methods, respectively. Mel-Frequency Cepstral Coefficients (MFCCs) were used as feature vectors for model training and speaker testing [9, 10]. 16 MFCCs were extracted at every 10 ms using 20 ms speech frame length. A Gaussian Mixture Model - Universal Background Model (GMM-UBM) speaker verification system [11] is used to perform experiments for male and female speakers separately. The gender-dependent background models were created using the concatenated speech of 120 male and 120 female speakers separately. Each UBM was constructed using 1024 mixtures. The claimant speaker models were derived from the gender-dependent UBMs using Bayesian adaptation. The number of claimant male and female speakers were 112 and 56, respectively. The verification score of a claimed speaker was determined by the log-likelihood ratio calculation. The results are reported as Equal Error Rate (EER) values. Table 1 and Table 2 show the EER values of the speaker verification system using different combinations of the training and the testing data for male and female speakers, respectively.

Row	Training Speech	Testing Speech	EER (%)
A	Uncoded	Uncoded	1.34
B	Uncoded	Org-LSF TGSM	3.59
C	Uncoded	NA-LSF TGSM	3.14
D	Org-LSF TGSM	Org-LSF TGSM	2.23
E	NA-LSF TGSM	NA-LSF TGSM	1.75
F	Org-LSF TGSM	NA-LSF TGSM	1.86
G	NA-LSF TGSM	Org-LSF TGSM	1.77

Table 1. EER values of verification system for male speech using TIMIT8k, Original-LSF TGSM, and NA-LSF TGSM databases.

Tables 1 and 2 show that the use of NA-LSFs in the GSM-EFR coder reduces the amount of loss in speaker verification performance. By employing NA-LSF extraction instead of the classical LSF extraction method in the GSM-EFR coder, the speaker verification EER values reduce from 3.59% to 3.14% and 2.23% to 1.75% for male speakers, and 6.25% to 5.39% and 2.52% to 1.83% for female speakers in the mismatched and matched training and testing conditions respectively. Experimental results also show that using NA-LSF coded speech in only training or the testing process (e.g. the training speech is collected from the original-LSF GSM-EFR coder and the testing speech is collected from the NA-LSF GSM-EFR coder) improves the speaker verification performance. The same performance improvement observed on

Row	Training Speech	Testing Speech	EER (%)
A	Uncoded	Uncoded	1.79
B	Uncoded	Org-LSF TGSM	6.25
C	Uncoded	NA-LSF TGSM	5.39
D	Org-LSF TGSM	Org-LSF TGSM	2.52
E	NA-LSF TGSM	NA-LSF TGSM	1.83
F	Org-LSF TGSM	NA-LSF TGSM	1.92
G	NA-LSF TGSM	Org-LSF TGSM	1.79

Table 2. EER values of verification system for female speech using TIMIT8k, Original-LSF TGSM, and NA-LSF TGSM databases.

whole of the Detection Error Trade-off (DET) curves. The performance increase is the result of using the NA-LSF parameter extraction method which removes the unwanted LSF track components in the frequency domain. More stable coefficients are obtained using the low-pass filtering operation, producing higher quality synthesised speech compared to the original LSF extraction of the GSM-EFR coder. As a result, the speaker verification performance on average is improved by 12.5% and 21.5% for male, and 13.8% and 27.4% for female speakers, under mismatched and matched conditions respectively. For female speakers, the EER value given in the row G of Table 2 is better than the EER value given in the row E. It is not clear why this particular EER value is smaller than the result of NA-LSF training/testing speech experiment. This result is currently being investigated. Although this method requires extra computational cost and time delay, as shown in [6] the method improves the synthesised speech quality, while providing easier quantisation compared to the original LSF extraction methods. Initial experiments indicate that the NA-LSF parameter extraction method can be used with any speech coder that employs LP analysis in its structure.

4. CONCLUSION

In this paper, the NA-LSF parameter extraction process for speaker recognition applications has been presented. It was shown that LSF vectors obtained with classical extraction methods contain undesired frequency components. These components cause some aliasing noise in the LSF parameters. The NA-LSF parameter extraction approach has been introduced in order to remove the undesired frequency components on the LSF tracks of the GSM-EFR coder. The speaker verification system experiments were performed using GSM-EFR coded speech in mismatched and matched conditions. The results obtained from these experiments show that the use of NA-LSF parameter extraction in the GSM-EFR coder increases speaker verification performance and by employing the NA-LSF method in the GSM-EFR coder reduces the speaker verification error by 12.5% and 21.5% for male, and 13.8% and 27.4% for female speakers under mismatched and matched conditions respectively. The proposed method is fully compatible with the existing standard speech coders, and thus it does not require any modification to existing infrastructures.

5. REFERENCES

[1] T.F. Quatieri, E. Singer, R.B. Dunn, D.A. Reynolds, and J.P. Campbell, "Speaker and Language Recognition using Speech Codec Parameters", *Eurospeech'99*, pp. 787-790, 1999.

[2] R.B. Dunn, T.F. Quatieri, D.A. Reynolds, and J.P. Campbell, "Speaker Recognition from Coded Speech and the Effects of Score Normalization", *35th Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 1562-1567, 2001.

[3] H. Altincay, C. Ergun, and W. Ahmad, "Robust Speaker Verification in Low Bit Rate Channels", *IEE Electronics Letters*, vol. 39, no. 6, pp. 576-578, 2003.

[4] ETSI, Recommendation GSM 06.60, "GSM Enhanced Full Rate Speech Transcoding". European Telecommunications Standards Institute, 1997.

[5] 3GPP, "AMR Speech Codec; General Description (Release 6)". 3rd Generation Partnership Project documentation, (3GPP TS 26.071 v6.0.0), 2004.

[6] K. Al-Naimi, S. Villette, and A. Kondoz, "Improved LSF Estimation Through Anti-Aliasing Filtering", *IEEE Speech Coding Workshop*, 2002.

[7] B. Atal, and M. Schroeder, "Adaptive Predictive Coding of Speech Signals", *The Bell System Technical Journal*, pp. 1973-1987, 1970.

[8] W.M. Fisher, G.R. Doddington, and K.M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status", *DARPA Workshop*, pp. 93-99, 1986.

[9] D.A. Reynolds. "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models". *In IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72-83, 1995.

[10] F. Bimbot, I. Magrin-Chagnolleau, and L. Mathan, "Second-order Statistical Measures for Text-Independent Speaker Identification". *Speech Communication*, vol. 17, no. 1-2, pp. 177-192, 1995.

[11] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Speech Processing*, vol. 10, pp. 19-41, 2000.