DETECTION OF OVERLAPPING SPEECH IN MEETING RECORDINGS USING THE MODIFIED EXPONENTIAL FITTING TEST

Angela Quinlan and Futoshi Asano

AIST Information Technology Research Institute,
Tsukuba, Ibaraki 305-8568,
JAPAN

ABSTRACT

Detection of overlapping speech in meeting recordings is a challenging problem due to both the nature of the conversation itself and the surrounding environment. Accurate identification of these sections of the recording is a crucial first step for speech recognition techniques as their non-detection leads to severe degradation in performance. A possible approach to solving this problem is the use of source number estimation techniques based on the ordered profile of the spatial correlation matrix eigenvalues.

In this paper we propose two approaches for detecting overlapping speech based on the Exponential Fitting Test (EFT) a source number estimation technique proposed in [1]. Firstly we propose a frequency domain implementation of the EFT, which is more appropriate when dealing with the broadband speech signals encountered in meetings. We then propose a second approach in which a correction factor is added to allow for the presence of reverberation. The performances of the proposed schemes are evaluated and compared to that of the original EFT using real meeting recordings.

1. INTRODUCTION

Analysis of meetings and multi-party conversations is reliant on accurate identification of overlapping segments of speech [2]. These segments occur frequently in natural conversations and result in severe degradation of automatic speech to text transcription. However, accurate detection of overlapping speech segments is a difficult problem due to both the nature of the speech and the environmental effects such as background noise, echo and reverberation [3]. Moreover, any suitable algorithm should be computationally simple, in order to allow for real-time implementation [4], and the amount of training data required should be minimal.

Previously suggested methods of detecting overlapping speech segments include the use of Support Vector Regression [5], which detects about 50% of the overlapping speech segments; and the extraction of acoustic features for use with a Gaussian Mixture Model (GMM) [3], which assumes that each participant has an individual microphone, and requires training of the classifier for each combination of features.

An alternative approach to detecting the presence of overlapping speech is to estimate the number of sources present. Any segment containing more than one source signal can then be classified as overlapping speech.

In some cases the second source present may not actually be a speech source and might instead be due to laughter or coughing. However, as these situations also lead to degradation of speech recognition techniques it is likewise important to identify such sections of the recording.

Classical source number determination techniques are based on eigen-decomposition of the observed spatial correlation matrix. Under ideal conditions (i.e. high Signal to Noise Ratio and a large number of uncorrelated data samples) the eigenvalues corresponding to the noise subspace are equal and the number of sources present is easily determined as the number of non-equal eigenvalues.

The most well known of these source number determination techniques are the Akaike Information Criterion (AIC) [6] and Rissanen's Minimum Description Length (MDL) [7]. However, due to the difficult operational conditions encountered in meeting recordings they are no longer accurate, as even at high Signal to Noise Ratio (SNR), they continuously over-estimate the number of sources present [1].

Recently an Exponential Fitting Test (EFT) was introduced for determining the number of speakers present in a mildly reverberant environment [1]. This method exploits the exponential profile of the ordered noise eigenvalues discussed in [8]. Under the assumption of white noise the profile of the noise-only eigenvalues is predicted using this exponential model together with the smallest eigenvalue, which is assumed to be a noise eigenvalue.

When the observed eigenvalues are compared to this predicted profile, the noise eigenvalues match the predicted values. Eigenvalues corresponding to the signal subspace are then easily identified as the values causing a break from the predicted profile.

In [1] this method is applied for controlled experiments with a fixed number of sources, for which case the EFT is shown to correctly determine the number of sources present with a probability of $55\% - 75\%$. However, we show here that under the operational conditions encountered in meetings the performance of the EFT is no longer adequate. In particular the presence of reverberation results in a large increase in the probability of false alarms, as discussed in section 6.

In this paper we therefore consider the situation where the number of sources must be estimated in the presence of reverberation. To this end we propose the use of a correction factor which corrects the predicted noise eigenvalue profile, in order to compensate for the presence of the reverberant tail of the first source signal. Once the presence of at least one speaker has been determined the observed eigenvalues are then compared to this corrected profile in order to distinguish between single and overlapping speech events.

2. PROBLEM FORMULATION

We consider the model of an array of M microphones located in a sound field generated by d sources, which are assumed to be non-coherent. Then, taking the short-term Fourier trans-

form of the signals received by the microphones, we obtain the following data model:

$$\mathbf{x}(\omega,T) = \mathbf{A}(\omega,T)\,\mathbf{s}(\omega,T) + \mathbf{n}(\omega,T), \quad (1)$$

where ω is the frequency under consideration, T is the frame index, $\mathbf{s}(\omega,T)$ is the source spectrum at time T and $\mathbf{A}(\omega,T)$ is the matrix of d direct path transfer function vectors. The spatial correlation matrix $\mathbf{R_{xx}}(\omega)$ is therefore defined as:

$$\mathbf{R_{xx}}(\omega) = E\left[\mathbf{x}(\omega,T)\,\mathbf{x}^H(\omega,T)\right] \quad (2)$$

Then defining $\mathbf{R}_{ss}(\omega)$ as the spatial correlation of the source signals; $\mathbf{I}$ as the $M \times M$ identity matrix; and assuming the noise $\mathbf{n}(\omega,T)$ is spatially white and uncorrelated from the sources with power σ^2; (2) can be re-expressed as:

$$\mathbf{R_{xx}}(\omega) = \mathbf{A}(\omega)\,\mathbf{R}_{ss}(\omega)\,\mathbf{A}^H(\omega) + \sigma^2(\omega)\,\mathbf{I}, \quad (3)$$

The eigenvalues of $\mathbf{R_{xx}}(\omega)$ are therefore given by:

$$\lambda_1(\omega),\ldots,\lambda_M(\omega) = \gamma_1(\omega)+\sigma^2(\omega),\ldots,\gamma_d(\omega)+\sigma^2(\omega), \sigma^2(\omega),\ldots,\sigma^2(\omega). \quad (4)$$

Assuming that the source power is greater than that of the background noise, the number of sources present can now be easily determined as the number of eigenvalues not equal to σ^2.

In practice however, $\mathbf{R_{xx}}(\omega)$ is unknown and an estimate is made using $\widehat{\mathbf{R}}_{\mathbf{xx}}(\omega) = \frac{1}{N}\sum_{T=1}^{N}\mathbf{x}(\omega,T)\,\mathbf{x}^H(\omega,T)$, where N is the number of frames the spatial correlation is taken over. The "signal eigenvalues" are still identified as the d largest ones. But, with the statistical fluctuations in $\mathbf{R_{xx}}(\omega)$, even in the presence of white noise the noise eigenvalues are no longer all equal to σ^2. In this case the separation between them is only clear in the case of high Signal to Noise Ratio (SNR) and low reverberation, when a gap can be clearly observed between signal and noise eigenvalues.

3. PROPOSED MODIFICATIONS OF THE EXPONENTIAL FITTING TEST

3.1 Frequency Exponential Fitting Test (FEFT)

Originally, the EFT was applied to the eigenvalues of the time-domain spatial correlation matrix [1]. However, as speech signals are broadband in nature the frequency domain spatial correlation matrix is used here as shown in equation (2), and the EFT is then applied at each individual frequency value of interest:

$$\omega = \omega_l : \omega_h, \quad (5)$$

where ω_l and ω_h are respectively the lowest and highest frequency values under consideration. The number of sources at each of these frequencies is determined and the individual results are then combined to give an overall decision on whether or not overlapping speech is present in the frame.

In order to determine the number of sources at each frequency the EFT predicts the decreasing profile of the eigenvalues of the noise spatial correlation matrix $\mathbf{R_{nn}}(\omega) = \frac{1}{N}\sum_{T=1}^{N}\mathbf{n}(\omega,T)\,\mathbf{n}^H(\omega,T)$, and compares the profile of the observed eigenvalues to this predicted profile.

As $\mathbf{R_{nn}}(\omega)$ has a Wishart distribution [8] it is extremely difficult, if not impossible, to find the decreasing profile of its eigenvalues. In the EFT this profile is instead approximated using the first and second order moments of the eigenvalues together with an initial assumption of white noise [9]. The smallest observed eigenvalue is assumed to be a noise eigenvalue, corresponding to a noise subspace dimension of $P = 1$. Then letting $P = P+1$ for each subsequent step until $P = M-1$, the predicted profile of the noise only eigenvalues is found recursively using:

$$\widehat{\lambda}_{M-P}(\omega) = (P+1)\,J_{P+1}\,\widehat{\sigma}(\omega)^2, \quad (6)$$

where:

$$J_{P+1} = \frac{1-r_{P+1,N}}{1-(r_{P+1,N})^{P+1}}; \quad (7)$$

$$\widehat{\sigma}(\omega)^2 = \frac{1}{P+1}\sum_{i=0}^{P}\lambda_{M-i}(\omega); \quad (8)$$

$$r = e^{-2a(M,N)}; \quad (9)$$

and:

$$a(M,N) = \sqrt{\frac{1}{2}\left\{\frac{15}{M^2+2} - \sqrt{\frac{225}{(M^2+2)^2} - \frac{180M}{N(M^2-1)(M^2+2)}}\right\}}. \quad (10)$$

The relative differences between the predicted and observed profiles can be found from:

$$r_m(\omega) = \frac{\lambda_m(\omega) - \hat{\lambda}_m(\omega)}{\hat{\lambda}_m(\omega)}, \qquad m = 1,\ldots,M-1, \quad (11)$$

and $r_m(\omega)$ is then compared to a threshold value $\eta_m(\omega)$ in order to determine whether or not a break from the noise-only profile has occurred.

The test described up to this point is a frequency domain implementation of the original EFT and in the following will be referred to as the Frequency Exponential Fitting Test (FEFT).

3.2 Corrected Frequency Exponential Fitting Test (CFEFT)

3.2.1 Application of the FEFT in the Presence of Reverberation

As with the EFT, the FEFT is based on the assumption that the background noise can be modelled as spatially and temporally white noise. This approximation is valid in many practical situations when there are no speakers present, making the FEFT suitable for the initial determination of whether or not there are any speakers present, in which case $r_1(\omega)$ will be less than the corresponding threshold $\eta_1(\omega)$.

However if $r_1(\omega) > \eta_1(\omega)$, then we know that there is at least one speaker present. The reverberant tail of this signal then leads to a violation of the initial assumption of white noise, leading to an increase in the noise eigenvalues relative to the predicted white noise values.

In this case the noise eigenvalue profile predicted from equations (6)-(10) will be lower than that of the observed

noise eigenvalues, resulting in frequent mis-classification of single speech segments as overlapping speech, which we call false alarms.

Therefore once it is known that speech is present in the signal the presence of the reverberant tail means that the white noise approximation no longer holds and it is necessary to apply a correction factor to the predicted profile in order to account for the increase in the noise eigenvalues due to reverberation.

The test resulting from the application of the correction factor to the predicted profile (at each frequency component) is called the Corrected Frequency Exponential Fitting Test (CFEFT) in the following.

3.2.2 *Calculation of the Correction Factor*

In order to calculate a suitable correction factor the eigenvalues of the estimated reverberation correlation matrix, $\lambda_1^{rev}(\omega),\ldots,\lambda_M^{rev}(\omega)$, are found. These values are then used to find the corresponding predicted noise eigenvalues $\hat{\lambda}_1^{rev}(\omega),\ldots,\hat{\lambda}_M^{rev}(\omega)$ as described in section 3.1.

The difference between the predicted and observed profiles, relative to the largest observed eigenvalue, is then taken as a correction factor:

$$cf_m(\omega)=\frac{\lambda_m^{rev}(\omega)-\hat{\lambda}_m^{rev}(\omega)}{\lambda_1^{rev}(\omega)},\qquad m=2,\ldots,M. \tag{12}$$

Once the presence of at least one source has been detected the correction factor is then used to modify the originally predicted noise eigenvalue profile:

$$\hat{\lambda}_m^{mod}(\omega)=(1+cf_m(\omega))\,\lambda_1(\omega). \tag{13}$$

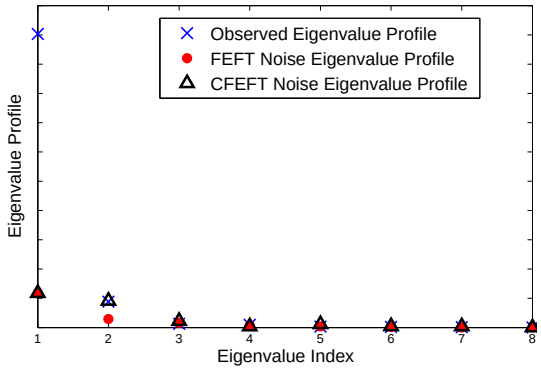
Figure 1 shows a comparison of the eigenvalues of the observed spatial correlation matrix, $\widehat{R}_{xx}(\omega)$ with the noise eigenvalue profiles predicted by the FEFT and the CFEFT.

In figure 1(a) we consider the case where a single speaker is present and therefore only the first observed eigenvalue, λ_1, should be greater than the corresponding predicted noise eigenvalue (it should be noted that the correction factor is only added to $\lambda_2,\ldots,\lambda_M$, not to λ_1). In the case of the FEFT the increase in the observed noise eigenvalues due to the presence of reverberation means the predicted noise profile is too low, leading to the incorrect detection of a 2nd source. However, it can be seen that the corrected profile of the CFEFT compensates for this increase and the predicted noise eigenvalues accurately model the observed noise eigenvalues allowing for correct determination of a single source.

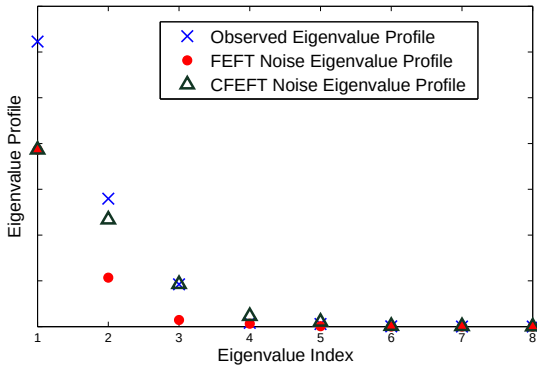
Figure 1(b) depicts the situation where two speakers are present In this case both λ_1 and λ_2 should be greater than the predicted profiles, as is the case for both the FEFT and the CFEFT. It can therefore be seen that while the correction factor increases the predicted noise profile, this increase does not mask the signal eigenvalues and detection of overlapping segments of the conversation is still possible.

Once again the predicted and observed profiles are compared by finding their relative difference:

$$r_m^{mod}(\omega)=\frac{\lambda_m(\omega)-\hat{\lambda}_m^{mod}(\omega)}{\hat{\lambda}_m^{mod}(\omega)}. \tag{14}$$



(a) Single speech segment



(b) Overlapping speech segment.

Fig. 1. Ordered Profiles of the observed and predicted eigenvalues.

It is important to note that the correction factor is found here from an estimate of the reverberation correlation matrix taken before the meeting begins (during a period with no speech) and is not updated during the meeting.

4. APPLICATION OF FEFT AND CFEFT

Based on the tests described in the previous section we now propose an algorithm for detection of overlapping speech segments. The proposed tests are identical except for the introduction of the correction factor in the CFEFT. Therefore letting $cf_m(\omega)=0$, for $m=2,\ldots,M$ in the FEFT and finding $cf_m(\omega)$ as described in section 3.2.2 for the CFEFT, the resulting algorithm is the same for both tests.

In this paper we are concerned with distinguishing between single and overlapping speech events, which is done by comparing $r_m(\omega)$ and $\eta_m(\omega)$ for $m=1,2$. The decisions made across the frequency range $\omega=\omega_l:\omega_h$ are then combined to produce an overall decision on whether or not multiple sources are present. The test is easily extended to determine the total number of sources present by comparing $r_m(\omega)$ and $\eta_m(\omega)$ for all $m=1,\ldots,M-1$.

4.1 Threshold Selection

The performance of the test is dependent on finding a suitable threshold η_m for $m=1,\ldots,M-1$. These threshold values are selected from the distribution of the relative difference for each frequency component when there is only noise present

at that frequency.

The initial threshold is calculated based on the first 750*ms* of the recording, which we assume are known to contain noise only. As a brief delay between the beginning of the recording and the beginning of the conversation is usual, this assumption is not thought to be restrictive. This initial estimate of the background noise is then updated during periods of silence for each frequency component throughout the meeting.

The choice of thresholds represents a compromise between the desired number of false alarms, i.e. the number of times that overlapping speech is mistakenly detected, and the rate of non-detection of overlapping speech. Depending on the proposed application of the test, errors due to false alarms may be more serious than non-detection errors, or vice versa.

Non-detection of overlapping speech segments can lead to severe degradation of subsequent speech recognition attempts. However, very high false alarm rates mean that large sections of the recording will be needlessly discarded, and the resulting lack of data may then make transcription of the meeting impossible.

The threshold $\eta_m(\omega)$ is found for $m = 1,\ldots,M$ and $\omega = \omega_l : \omega_h$ as the value greater than a pre-defined percentage of the Q previous noise-only relative difference values, $r_m(\omega)$. In the following we call this percentage the "threshold step". The value of Q corresponds to the "memory" of the algorithm and can be increased or decreased depending on the variation of the background noise. In this case we use $Q = 50$.

4.2 Silence Detection

In order to avoid the propagation of errors due to poor threshold selection, the data blocks used to update the threshold are selected independently of their eigenvalue distribution. Instead the noise estimate for each frequency is found by comparing the corresponding energy to the energy threshold for that frequency [10]. This energy threshold is updated for each block and is given by:

$$\psi(\omega,k) = \beta E(\omega,k-1); \tag{15}$$

where k is the block index, (with N frames in a block). $E(\omega,k-1)$ is the energy of the previous noise at the given frequency, and β is a constant value lying between 1.5 and 2.5, which in this case is equal to 1.7.

5. EXPERIMENTAL SETUP

The EFT (as proposed in [1]), the FEFT and the CFEFT were then tested using recordings of a Japanese market research meeting where one interviewer and five interviewees were positioned around a table. Throughout the meeting the interviewer asked questions which the interviewees then responded to in a discussion-type manner.

The meeting was conducted in a middle sized meeting room with a reverberation time of 500*ms*. A circular microphone array with a diameter of 15*cm* and consisting of 8 microphones was placed in the middle of the table and the distance from the centre of the array to the participants was approximately $1.0-1.5m$, allowing for the assumption of far-field sources to be made.

The recorded signals are broken into overlapping blocks of length 500*ms*, and for processing in the frequency domain these blocks are further divided into frames of length 32*ms*.

Sampling Frequency	16000Hz
FFT Length	512
FFT Shift	128
Frequency Range	500-4000Hz
Block Length	0.5s
Block Overlap	0.25s

Table 1. Experimental Parameters

The spatial correlation value for each block is then found by averaging across the results from each frame. A full list of the experimental parameters used is given in table 1.

6. RESULTS

The performance of the three tests was compared based on the calculation of the Recall Rate R_r; the Precision Rate R_p; the False Alarm Rate R_f; and the F-measure; as defined below [5]:

$$R_r = \frac{\text{Number of Correctly Detected Overlaps}}{\text{Total Number of Overlaps Present}}, \tag{16}$$

$$R_p = \frac{\text{Number of Correctly Detected Overlaps}}{\text{Total Number of Overlaps Detected}}, \tag{17}$$

$$R_f = \frac{\text{Number of Incorrectly Detected Overlaps}}{\text{Total Number of Overlaps Detected}}, \tag{18}$$

$$F = \frac{2R_rR_p}{R_r+R_p}. \tag{19}$$

As the F-measure is the harmonic mean of the precision and recall, this score can be taken as the most important measure of each test.

The performance of each of the tests is dependent on the selection of a suitable threshold step, and therefore we compare the evolution of the F-measures of the three tests across the range of possible threshold steps in order to find the best threshold step for each test.

From the results shown in figure 2 it can be seen that the CFEFT offers a significant improvement in the maximum F-measure that can be achieved. Using these optimum thresholds the corresponding R_r, R_p and R_f values are then reported in table 2. It can be clearly seen that the CFEFT offers the best overall performance. The CFEFT results also show a significant improvement compared to those previously reported in literature [5].

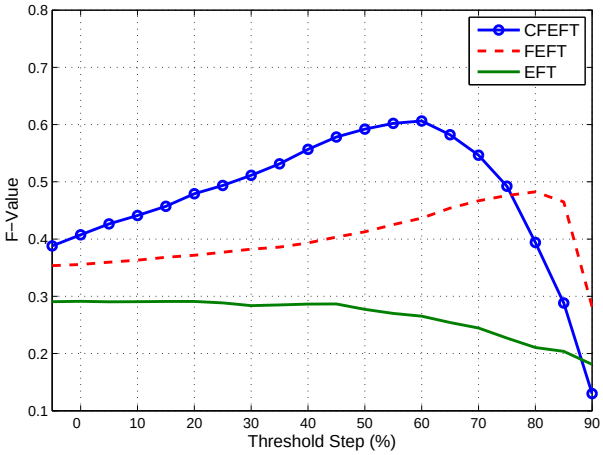


Fig. 2. F-measure for the CFEFT, FEFT and EFT as the threshold step is increased.

	Rr	*Rp*	*Rf*	**F**
EFT	0.84	0.18	0.82	**0.29**
FEFT	0.71	0.36	0.64	**0.48**
CFEFT	0.69	0.54	0.46	**0.61**

Table 2. Results

The correction factor used here for the CFEFT is based on an initial estimate of the reverberation present. As the spatial correlation of the reverberation changes throughout a meeting, updating this estimate of the reverberation would be expected to result in improved performance of the test.

Analysis of the results indicates that false alarms occur in blocks where there is a change over between speakers and there is a pause before the next speaker begins. This may be due to the presence of some remaining speech or noise signal from the previous speaker.

7. CONCLUSION

In this paper we present two methods for detecting overlapping speech in meeting recordings based on the Exponential Fitting Test (EFT), a model order determination test proposed in [1].

Firstly we proposed a frequency domain implementation of the EFT, the Frequency Exponential Fitting Test (FEFT), which results in a dramatic increase in the probability of correctly detecting overlapping speech. However this approach also results in an increase in the probability of false alarm, making it unsuitable for practical applications.

The second test proposed here, the Corrected Frequency Exponential Fitting Test (CFEFT), is similar to the FEFT. However, in this case a correction factor is introduced to allow for the effects of reverberation. From the results it can be seen that this approach offers a significant increase in performance compared to both the other tests and previously reported results.

8. ACKNOWLEDGMENTS

This work is support by a Japanese Society for the Promotion of Science (JSPS) postdoctoral fellowship.

REFERENCES

[1] A. Quinlan, F. Boland, J. Barbot, and P. Larzabal, "Determining the Number of Speakers with a Limited Number of Samples," in *European Signal Processing Conference EUSIPCO*, Florence, 2006.

[2] F. Asano and J. Ogata, "Detection and Separation of Speech Events in Meeting Recordings," in *Proc. 9th International Conference on Spoken Language Processing (INTERSPEECH)*, Pittsburgh, PA, 2006.

[3] S. Wrigley, G. Brown, V. Wan, and S. Renals, "Speech and Crosstalk Detection in Multichannel Audio," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 1, pp. 84–91, 2005.

[4] P. D. Leon, "Short-time Kurtosis of Speech Signals with Application to Co-channel Speech Separation," in *Multimedia and Expo, 2000, ICME 2000*, New York, NY, 2000.

[5] K. Ymamoto, F. Asano, T. Yamada, and N. Kitawaki, "Detection of Overlapping Speech in Meetings Using Support Vector Regression," in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Eindhoven, The Netherlands, 2005.

[6] H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Trans. Automat. Contr.*, vol. 19, no. 6, pp. 716–723, 1974.

[7] J. Rissanen, "Modelling by Shortest Data Description Length," *Automatica*, vol. 14, pp. 465–471, 1978.

[8] J. Grouffaud, P. Larzabal, and H. Clergeot, "Some Properties of Ordered Eigenvalues of a Wishart Matrix: Application in Detection Test and Model Order Selection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Atlanta, GA, 1996.

[9] A. Quinlan, J.-P. Barbot, P. Larzabal, and M. Haardt, "Model Order Selection for Short Data: An Exponential Fitting Test (EFT)," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, p. Article ID 71953, 2007.

[10] H. Hirsch and C. Ehrlicher, "Noise Estimation Techniques for Robust Speech Recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Detroit, MI, 1995.