Environmental Robust Features for Speech Detection

Josep M. Solà i Carós, Yin H. Lam and Thomas Kemp
European Technology Center
Sony International (Europe) GmbH
jmsola@ieee.org lam@sony.de kemp@sony.de

Climent Nadeu
TALP Research Center
Universitat Politècnica de Catalunya (UPC)
climent@talp.upc.es

Abstract

In this paper, two novel features, Line Spectrum Center Range and Line Spectrum Flux, both derived from Line Spectrum Frequencies, are proposed to detect the presence of speech in various acoustic environments. Evaluation results using Fischer Discriminant Analysis and Scatter Matrices indicated that the new features excel the state-of-the-art features. An environmental robust hybrid feature set including the proposed features, Normalized Energy Dynamic Range and Mel-Frequency Cepstrum Coefficients is further introduced. When evaluating the hybrid feature set on a Gaussian Mixture Model based classification engine, the results showed that the hybrid feature set outperformed Mel-Frequency Cepstrum Coefficients up to 49% in terms of relative frame error rate.

1. Introduction

Robust speech detection algorithms are essential to many applications such as personal hearing aid systems in which robust speech/noise discrimination is a pre-requisite for effective noise reduction [1], or speech recognition systems where poor speech detection usually lead to significant degradation in recognition accuracy [2]. Conventional energy based speech detection systems work well in clean environments. However, they fail to handle the complex acoustic situations encountered in many real life applications where the signal energy level is usually highly dynamic and background sounds such as music and non-stationary noise bursts are common. Alternatively, model based speech segmentation algorithms were widely introduced to provide a reliable solution to distinguish speech from other complex environment sounds. Features such as full band energy, sub-band energy, linear prediction residual energy, pitch, zero crossing rate related features and centroid etc., are usually employed in such systems [2] [3]. However, they are either energy dependent or not robust against the environment changes. Thus, model re-training is usually required in order to port the system to a different application domain. On the other hand, although numerous features have been proposed in the literature, no clear conclusion about the optimal feature set has ever been reached for speech detection tasks.

In this paper, two novel features derived from Line Spectrum Frequencies (LSF) [4], which are robust against changes in acoustic environments, are introduced. Two feature selection algorithms, Fischer Discriminant Analysis and Scatter Matrices, were used to compare the proposed features to other state-of-the-art features. In addition, the performance of Line Spectrum Frequencies is compared to the state-of-the-art Mel-Frequency Cepstrum Coefficients [5] [6]. Finally, an optimal hybrid feature set is proposed and its performance on home environment data and TV broadcasting data, two completely different application scenarios, is investigated.

The rest of the paper is organized as follows: Section 2 gives a review of the state-of-the-art features, followed by the proposed LSF derived features. Section 3 describes the experimental setup. Evaluation results and discussion are depicted in Section 4 and finally Section 5 gives the conclusions.

2. Feature Extraction

2.1. State-of-the-art Features

In order to efficiently detect the presence of speech in various acoustic environments, only those features that are energy independent and that have been proved to give robust performance are selected. To mention:

- **Spectral Centroid (SC):** It gives a measure of the Center of Gravity of the spectral energy distribution. While voiced speech tends to have low SC values, unvoiced speech gives high SC values. On the other hand, because of the energy’s distribution of music, centered SCs are frequently observed [7].

- **Spectral Centroid Range (SCR):** Because speech presents strong alternations between voiced and unvoiced events, SC ranges of speech signal computed over a temporal window tend to have a higher value when compared to non-speech signals.

- **Spectral Flux (SF):** SF gives the average spectral amplitude variation between two adjacent frames within a temporal window. The SF values of speech are usually higher than those of music and stationary noise [7].

- **Spectral Rolloff (SR):** It is defined as the 95th percentile of the power spectral distribution. The physical
interpretation is similar to \(SC\) [7].

- **High Zero Crossing Rate Ratio (HZCRR):** It is defined as the ratio of frames whose ZCR is higher than 1.5-fold \(\overline{ZCR}\) computed over a temporal window. HZCRRs of speech are expected to be higher than non-speech due to the voiced-unvoiced alternating nature of the speech signals [3].

- **Normalized Energy Dynamic Range (NEDR):** It is defined as the range of the temporal energy normalized by the maximum energy within a time window [8].

### 2.2. Line Spectrum Frequencies (LSF)

Line Spectrum Frequencies is an alternative representation of Linear Prediction Coefficients (LPC). Because of its well-behaved dynamic range and also their filter stability preservation property [4], it has been commonly used in speech coding applications. From the all-zero inverse filter \(A(z)\) in the LPC analysis, a symmetric polynomial \(P(z)\) and an anti-symmetric polynomial \(Q(z)\) can be defined. Such transformation map each complex root of \(A(z)\) to a root in \(P(z)\) and \(Q(z)\) respectively. The roots of \(P(z)\) and \(Q(z)\) are interlaced with each other and they lie on the unit circle, i.e., roots of \(P(z)\) and \(Q(z)\) are described only by their phases \(\omega_P\) and \(\omega_Q\). \(\omega_P\) and \(\omega_Q\) hence define the so called Line Spectrum Frequencies.

Detailed examination of the LSF and LPC frequency interpretation reveals the following interesting properties: the difference between each pair of \(\omega_P\) and \(\omega_Q\) is related to the prominence of the spectral peak described by the corresponding root of \(A(z)\). The smaller the distance, the sharper is the spectral peak. The average of \(\omega_P\) and \(\omega_Q\), on the other hand, provides the information about the location of the spectral peak. Therefore, the location and the prominence of spectral peak described by the complex root of \(A(z)\) are preserved in the LSF. Figure 1 shows the LPC reconstructed spectral envelope of a voiced speech frame extract from The Matrix Soundtrack. The relationship between LPC and LSF analysis is illustrated in Table 1. As observed, the minimum of \(\omega_P - \omega_Q\) occurs when \(p = 1\), and this correspond to the 2nd spectral peak which is the most prominent and the sharpest. On the other hand, the phase of a \(A(z)\) root, which estimates the spectral peak position, is approximately equal to \(\frac{\omega_P + \omega_Q}{2}\).

### 2.3. LSF Derived Features

Line Spectrum Frequencies (LSF) and differential Line Spectrum Frequencies (DLSF), the successive difference between \(\omega_P\) and \(\omega_Q\), have been shown to give satisfactory performance in the field of audio content analysis [5]. In this section, a novel feature, Line Spectrum Center Range (LSCR), which exploits both the voiced-unvoiced alternating behavior of speech and the robustness nature of LSF is first described. Afterwards, Line Spectrum Flux (LSFL), a new distance function which measures the average variation of LSF between two adjacent frames within a temporal window, is proposed.

#### 2.3.1. Line Spectrum Center Range

As depicted in Section 2.2, \(\frac{\omega_P + \omega_Q}{2}\) gives a good estimation of the position of a spectral peak. Here, a Line Spectrum Center \(LSC(f)\) at frame \(f\) is first introduced,

\[
LSC(f) = \sum_{p=0}^{P} MLSF^f(p) \tag{1}
\]

where \(MLSF^f(p) = \frac{\omega_P^f(p) + \omega_Q^f(p)}{2}\). Unlike Spectral Centroid, whose value is computed over the full signal spectrum and hence easily prone to the change of signal and noise levels in the application environments, \(LSC\) is computed using only the prominent spectral peak locations which are less likely to be influenced by the environment changes. Similar to \(SC\), \(LSC\) also differs significantly for voiced and unvoiced speech. Hence a Line Spectrum Center Range \(LSCR(f)\) at frame \(f\) is defined as,

\[
LSCR(f) = \max_{p} LSC(f^p) - \min_{p} LSC(f^p) \tag{2}
\]

Where \(f^p\) is the frame index within a temporal window centered at frame \(f\). Figure 2 displays the probability density distributions of \(LSCR\) and \(SCR\) for speech, music and noise. As observed, the overlap among speech and non-speech of \(LSCR\) is significantly lower than that of \(SCR\).
2.3.2. Line Spectrum Flux

As mentioned, the changes of spectral peak location in successive frames differs significantly between speech and non-speech. Also, inspired by the average inter-frame LSF distance (IFLSFD) proposed in [4], our proposed distance function LSFL, which provides a quantity measure of the average variation of the spectral peak locations between two adjacent frames within a temporal window of $F$ frames centered at frame $f$ is defined as,

$$LSFL(f) = \frac{1}{FF} \sum_{f' = 1}^{F} \sum_{p = 1}^{P} \left[ 1 - \frac{MLSF^f(p) - MLSF^{f'}(p-1)}{MLSF^{f'}(p) - MLSF^{f'}(p-1)} \right]$$

The probability density distributions of speech, music and noise of LSFL and IFLSFD are shown in Figure 3. It is observed that, on average, speech presents higher LSC values than those of the other audio events. When compared to IFLSFD, LSFL significantly improves the separation between speech and non-speech events. In addition, due to the quantization introduced in the implementation of Line Spectrum Frequencies, IFLSFD is occasionally numerical unstable. The numerical stability of LSFL is however guaranteed.

### Table 2: Fischer Discrimination Analysis.

<table>
<thead>
<tr>
<th>Feature</th>
<th>F-Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSCR</td>
<td>2.6</td>
</tr>
<tr>
<td>LSFL</td>
<td>2.2</td>
</tr>
<tr>
<td>SCR</td>
<td>1.7</td>
</tr>
<tr>
<td>NEDR</td>
<td>1.5</td>
</tr>
<tr>
<td>HZCRR</td>
<td>1.1</td>
</tr>
<tr>
<td>SR</td>
<td>0.5</td>
</tr>
<tr>
<td>IFLSFD</td>
<td>0.1</td>
</tr>
<tr>
<td>SC</td>
<td>0.1</td>
</tr>
<tr>
<td>SF</td>
<td>0.0</td>
</tr>
</tbody>
</table>

3. Experimental Setup

Two independent database were used in our experiments: a 7 hours TV database which consists of both German and English TV broadcasting material and a 20 hours Home database in which English speech was mixed with different real home recordings. Each database was split into a training and testing subsets. The TV test set is about 1 hour and the Home test set is around 1.5 hours. The format of the data is 16KHz sampling rate, mono and 16 bit per sample. For feature preprocessing, 32ms frame length, 10ms frame shift, 0.5s window length, a 512-pt FFT and 18th order LPC were used. For classification, Gaussian Mixture Models with 64 mixture components and diagonal covariance matrices were used to model the nature of speech and non-speech. In the evaluation, a maximum duration constraint of 250ms was incorporated into the Viterbi search by using a 25 states HMM engine.

4. Evaluation Results and Discussion

4.1. Feature Selection

Fisher Discriminant Ratio (F-Ratio) measures the ability of a feature to separate the underlying classes by taking the distances among the class and also the variances within individual classes into account [9]. Table 2 gives the F-Ratios extracted from the Home training database. As observed, LSCR, LSFL, SCR, NEDR and HZCRR demonstrated good performance in discriminating speech from non-speech. However, as F-Ratio does not consider the inter-correlation among features, Scatter Matrices which measures the ratio between within class scatter and total class scatter of a feature set [9], was used to select the optimal feature subset. According to the results presented in Table 3, the optimal feature set comprises LSCR, LSFL and NEDR. Because of the correlation between LSCR and SCR, SCR was not included in the optimal feature set. By combining the analysis results from F-Ratio and Scatter Matrices, it is concluded that $[LSCR, LSFL, NEDR]$ is the best feature set.
4.2. Evaluation Results

Table 4 shows the frame classification error rate for the Home test set using models trained on Home and TV training sets respectively. By considering the average performance in both matched (Home) and unmatched (TV) training conditions, MFCC without energy coefficient ($nemo_mfcc$) outperformed $LSF$ and $DLSF$ significantly. It is also noted that the addition of $\Delta$ and $\Delta\Delta$ in general degrades the performance. By combining these results with the conclusions from Section 4.1, a novel hybrid feature set is proposed as: [$nemo_mfcc, LSCR, LSFL, NEDR$].

The evaluation results for both Home and TV test data sets using models trained on either Home or TV data sets are displayed in Table 5. When comparing to the evaluation results showed in Table 6 using the baseline [6] $nemo_mfcc$ feature only, the relative error rates reduction for matched and unmatched environments are, in all cases, equal to or larger than 35%.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Match</th>
<th>Unmatch</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>$nemo_mfcc$</td>
<td>14.4%</td>
<td>17.8%</td>
<td>16.1%</td>
</tr>
<tr>
<td>$nemo_mfcc + \Delta + \Delta\Delta$</td>
<td>14.3%</td>
<td>21.5%</td>
<td>17.9%</td>
</tr>
<tr>
<td>$LSF$</td>
<td>21.4%</td>
<td>25.9%</td>
<td>24.6%</td>
</tr>
<tr>
<td>$DLSF$</td>
<td>17.5%</td>
<td>39.1%</td>
<td>28.3%</td>
</tr>
</tbody>
</table>

Table 5: The frame error rates of the hybrid feature set.

Table 6: The frame error rates of Mel-Frequency Cepstrum Coefficients without energy coefficient.

As the addition of LSF derived features increases the computation complexity of the feature set, in the future, the reduction of the computation complexity will be investigated. Since the features of the hybrid feature set might correlate among each other, further optimization of the classification engine might improve the results significantly: PCA and full covariance matrices are aimed to be introduced.

5. Conclusions and Future Work

In this paper, two novel LSF derived features which reliably detect the presence of speech in various acoustic environments are proposed. Used together with MFCC and Normalized Energy Dynamic Rate, an average relative improvement of 49% is observed when compared to using MFCC only [6].

References


