Comparison of Spectral Derivative Parameters for Robust Speech Recognition

Dušan Macho and Climent Nadeu

TALP Research Center
Universitat Politècnica de Catalunya, Barcelona, Spain
Email: {dusan,climent}@talp.upc.es

Abstract

Recently, spectral first-derivative parameters obtained by frequency filtering (FF) have been successfully used in both clean and noisy HMM speech recognition. In this paper, two types of spectral derivative parameters, the usual FF features and the relative spectral difference (RSD) features, are compared both between them and with their second-derivative versions. Additionally, another kind of recently introduced robust speech features, the SBCOR parameters, are related theoretically with the second-derivative RSD. By experimentally comparing all those types of features in the Aurora 2.0 noisy database framework, we conclude that the first-derivative parameters are preferable to the second-derivative ones (and to the MFCC) for both clean and noisy speech recognition, and the RSD parameters show the best average performance.

1. Introduction

Speech parameters employed in current ASR systems model the time evolution of the spectral envelope of the speech signal, i.e. a smoothed version of the speech spectrum that is usually obtained by filter-bank (FB) or linear prediction (LP) analysis.

Recently, spectral first-derivative parameters, obtained by filtering along the frequency variable the sequence of spectral energies [1], have been successfully used in both clean and noisy HMM speech recognition [2]. Actually, Klatt already showed that local slopes of the spectral envelope near the peaks play an important role in speech perception, more for instance than the FF energies (FBEs) [3]. Furthermore, the spectral first-derivative parameters show an interesting characteristic, which is absent in the mel-frequency cepstral coefficients (MFCCs): they are localized in frequency.

Actually, filtering along the frequency variable may be carried out either after or before applying the log non-linearity on the FB energies [2]. In the latter case, each filter output must be divided by the energy itself. Additionally, when filtering is a discrete differentiation, the parameters resulting from the later case, which were already introduced in [4], actually are relative spectral differences (RSD).

In this paper, we will assume that the spectral energies come from a FB analysis, and we will consider not only the spectral first-derivative parameters but also the spectral second-derivative ones. In fact, as shown in this paper, the second-derivative RSD is theoretically similar to another recently introduced speech representation, the SBCOR parameters [4].

All those speech parameterizations will be empirically compared in clean and noisy connected-digit tests in the framework of the Aurora 2.0 database and recognition backend [7], using only clean speech for training.

2. Log-spectral differences and relative spectral differences

Let us denote by $E(\omega)$ the speech spectral envelope of the current frame and $S(\omega)$ its log-compressed version, i.e.

$$S(\omega) = \log(E(\omega)).$$

(1)

The derivative of $S(\omega)$ can be written

$$\frac{d}{d\omega} S(\omega) = \frac{d}{d\omega} \log(E(\omega)) = \frac{1}{E(\omega)} \frac{d}{d\omega} E(\omega).$$

(2)

We herewith refer to the calculation of the log-spectral derivative as expressed by the right side of (2) as relative spectral derivative. Note that the log operator disappears and the differentiation is applied on the spectral envelope in this case. Thus, in the case of continuous frequency variable, the spectral slope of the log spectrum can be obtained either from the log-spectral derivative or from the relative spectral derivative.

When spectral energies are available at discrete frequency values (e.g. when a FB analysis is employed), the derivatives involved in (2) should be replaced by differences. Therefore, two techniques for the computation of local spectral slope measurements can be derived depending on the use of either the left or the right side of (2). By measuring the slope as the difference between the two samples surrounding the current one, the log-spectral derivative results in the log-spectral difference or frequency filtering (FF) parameter [1]

$$S_{FF}(k) = S(k + 1) - S(k - 1),$$

(3)

where $k$ is the frequency sub-band index, and the relative spectral derivative results in the relative spectral difference (RSD)

$$S_{RSD}(k) = \frac{E(k + 1) - E(k - 1)}{E(k)}.$$

(4)

Note that the FF parameters in (3) actually are the most usual type of frequency-filtered energies, which are obtained with the filter $z^{-1}$ [1]. In Figure 1, we graphically compare both FF and RSD parameters. Figure 1(a) shows the FB energies of the phoneme /e/, corresponding to 20 mel-spaced sub-bands of a speech signal from the Aurora 2.0 database. The corresponding FF and RSD parameters are displayed in the upper part of Figure 1(b). It can be observed that FF parameters show values close to zero at spectral peaks and valleys, and positive/negative values at increasing/decreasing local spectral slopes, respectively. The RSD parameters are close to the FF parameters at the spectral peaks and valleys; however they show higher absolute values at the positions of sharp spectral
slopes, due to a relatively low denominator value in (4). To avoid that behavior of the RSD parameters, we smoothed the denominator of (4) by using an average of all the sub-band energies included in the RSD computation, i.e.

\[
(5)
\]

As it can be observed from the lower part of Figure 1(b), the resulting RSD parameters resemble much more the FF parameters.

3. Relation between the second-derivative RSD and the SBCOR parameters

In [5], the authors proposed a new robust feature extraction technique called sub-band autocorrelation (SBCOR), which extracts the periodicities associated with the inverse of the center frequency of the analyzed sub-band. In this section, we will show the similarity between the SBCOR parameters and the second-derivative RSD parameters.

The sub-band energies \( E(k) \) used in the RSD computation (4) are obtained by integrating the estimated power spectrum \( P(\omega) \) (the periodogram) of the current frame within each sub-band as

\[
E(k) = \int_0^\pi |H_k(\omega)|^2 P(\omega) d\omega / \pi ,
\]

where \( H_k(\omega) \) is the transfer function of \( k \)-th sub-band filter. Thus, the numerator in (4) can be written as the following weighted integration of the spectrum

\[
E(k + 1) - E(k - 1) = \int_0^\pi W_k(\omega) P(\omega) d\omega / \pi
\]

where the weighting function is

\[
W_k(\omega) = |H_{k,1}(\omega)|^2 - |H_{k,-1}(\omega)|^2 .
\]

Figure 2(a) displays \( W_k(\omega) \) for the sub-band centered at frequency 1000 Hz when using the typical triangular mel-spaced filter bank.

The second difference may be obtained by convolving the first difference with itself. In this case, the numerator of (4) would be

\[
-2E(k) + E(k - 1) + E(k + 1) .
\]

The corresponding weighting function \( W_k(\omega) \) is depicted in Figure 2(b).

On the other hand, the SBCOR parameters are computed as

\[
S_{\text{SBCOR}}(k) = R_k(m_k) / R_k(0) , \quad m_k = \frac{2\pi}{\omega_k}
\]

with \( R_k(m) \) the autocorrelation function of the \( k \)-th sub-band, and \( \omega_k \) is the center frequency of that sub-band.
filter bank consisting of Q Gaussian filters defined as

$$|H_k(\omega)|^2 = e^{-2C_k(\omega-a_k)^2}, \quad \omega \geq 0$$

with

$$C_k = \frac{Q}{2} \ln 2 |\omega|^2$$

was considered in [5], where the FB center frequencies are equally spaced on the Bark scale.

The SBCOR parameters computed with the expression (10) can be interpreted as a kind of relative spectral differences. In fact, the denominator in (10) is a measure of the energy within the k-th sub-band, like the denominator of the RSD computation (4), and the numerator \( R_k(m_k) \) can be written as

$$R_k(m_k) = \frac{\pi}{0} W_k(\omega) |H_k(\omega)|^2 \cos(\omega m_k) d\omega / \pi$$

where

$$W_k(\omega) = |H_k(\omega)|^2 \cos(\omega m_k).$$

Figure 3 shows the shape of the weighting function using the Gaussian sub-band filter centered at 1105 Hz (the frequency axis is normalized by the sub-band center frequency). In [6], the authors realized the presence of the lateral inhibitive weighting in the computation of \( R_k(m_k) \) that can be noticed in this figure. From tests with a DTW word recognizer, they experimentally found worse results in recognition of noisy speech by using a weighting function like \( W_k(\omega) \) but suppressing its negative lobes; actually, the resulting weighting function corresponds to a zero-order derivative. Notice the similarity between the shapes of the weighting functions from Figure 3 and Figure 2(b). Therefore, the SBCOR parameters are similar to the second-derivative RSD parameters.

In total, eight real noises were used (e.g., noises from street, train, car, or babble noise).

The recognition system, which is the one used for the above mentioned standardization work, is based on continuous density HMMs with diagonal covariance matrices. Each digit is modeled by 18 states with 3 Gaussians per state and silence is modeled by 6 states with 6 Gaussians per state. Training was only based on clean speech in our experiments.

The recognition performance of each of the techniques considered in this paper is compared to that of the clean speech standard front-end [8]. In this front-end, 12 mel-frequency cepstral coefficients (MFCC) and the log energy coefficient are computed each 10ms from signal frames of length 25ms. Delta and acceleration coefficients are appended to this static set of features.

4.2. Implementation of FF, RSD and SBCOR

FF, RSD and SBCOR parameters are computed every 10ms from Hamming-windowed 30ms long signal frames. For the first-derivative FF and RSD parameters, 14 FB energies (FBE) were computed for each frame by integrating the FFT power spectrum according to mel-scale-distributed triangular weighting functions with 50% overlapping. Static FF parameters were computed with expression (3). Note that the two endpoints of each FF vector actually are absolute log FBE of the 2\textsuperscript{nd} and 13\textsuperscript{th} FB bands, \( S(2) \) and \( S(13) \). Static RSD parameters were computed by using (5). Both endpoints of each RSD vector, \( S_{RSD}(1) \) and \( S_{RSD}(14) \), were replaced by the absolute log FBE, \( S(2) \) and \( S(13) \) like in the FF case.

To compute the second-derivative FF and RSD parameters, we used 16 Mel-spaced FBE. In the FF case, the following expression was used

$$S_{2FF}(k) = -S(k+2) + 2S(k) - S(k-2).$$

and, in the case of RSD parameters, we used

$$S_{2RSD}(k) = \frac{-E(k+2) + 2E(k) - E(k-2)}{E(k+1) + E(k) + E(k-1)}.$$
4.3. Recognition performances

The set of three tests from the Aurora 2.0 database has been used to evaluate the performance of all the considered techniques. In each of the tests A and B, testing data were distorted by four different additive real noises at various SNR levels. In test C, besides the additive distortion, channel distortion is also included. Recognition rates from SNRs ranging from 20dB to 0dB were averaged. Table 1 summarizes them in terms of relative average improvements with respect to the standard MFCC parameterization.

From the first two lines of Table 1, we can observe that both first-derivative FF and RSD features obtain significant relative average improvements with respect to MFCC in recognition of both clean and real noise distorted digits. Robustness of FF features has already been reported elsewhere (see [2]). The relative improvements obtained by using RSD are slightly higher than those of FF features for noisy speech, but a significant difference can be observed for clean digits.

Notice that for test C, results for noisy speech are not so different from the MFCC’s ones. In this case, the artificial channel distortion actually is a filtering operation that reduces the additive noise distortion. Let us also note that we do not use any technique to cancel the linear distortion in these tests.

On the other hand, we did not observe consistent improvements when using the second-derivative FF or RSD parameters, but lower clean and channel-distorted digit rates than those of the MFCC have been obtained (see the negative values of the relative average improvements at lines 3 and 4 of Table 1). Actually, the second-derivative can enhance the spectral peaks but it is not a slope measure. From another viewpoint, the second derivative emphasizes higher quefrency indexes than the first derivative.

Similarly to the case of the second-derivative FF or RSD parameters, the clean speech recognition performance of SBCOR parameters is low (see line 5 in Table 1). In our implementation of SBCOR, we did not observe any significant improvement in noisy digit recognition over the MFCC standard parameterization.

The last two lines of Table 1 show the relative average improvements of the first-derivative FF and RSD parameters when combined with a linear spectral subtraction (LSS) pre-processing. The significant additional improvements in noisy digit recognition rates that are obtained in this way indicate that the performance improvement from LSS is complementary to that from FF/RSD. Note that even the clean digit rates have been further improved.

5. Conclusions

In this paper, we analytically and experimentally compared three techniques for robust feature extraction that differ from the usual MFCC scheme in the sense that the parameters, which are various types of spectral derivatives, are located in the frequency domain. The first-derivative FF and RSD parameters yielded significantly better noisy and clean digit recognition rates than the MFCC: in average, 30.89% relative average improvement in additive noise conditions (tests A and B, not including the convolutive noise test C). RSD showed a better performance than FF, specially for clean speech. An additional improvement has been obtained when combining FF and RSD with LSS pre-processing (41.06%). On the other hand, we did not observe any consistent advantage of the second-derivative-based features (FF, RSD, and SBCOR) over the MFCC standard technique in the performed tests.

6. Acknowledgments

The authors would like to thank to F. Galindo, J. Hernando, J. Mari, J. Padrell and J. Salavedra for helpful discussions. J. Mari carried out an early implementation of RSD, and F. Galindo developed the LSS. This work has been partially supported by Motorola Ltd. (UK), and also by CICYT under projects TIC98-0683 and TIC98-0423-C06-01.

7. References


<table>
<thead>
<tr>
<th>Technique</th>
<th>Test A</th>
<th>Test B</th>
<th>Test C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>clean</td>
<td>noisy</td>
<td>clean</td>
</tr>
<tr>
<td>1</td>
<td>1° derivative FF</td>
<td>8.67</td>
<td>24.24</td>
</tr>
<tr>
<td>2</td>
<td>1° derivative RSD</td>
<td>17.09</td>
<td>26.39</td>
</tr>
<tr>
<td>3</td>
<td>2° derivative FF</td>
<td>-36.99</td>
<td>-2.71</td>
</tr>
<tr>
<td>4</td>
<td>2° derivative RSD</td>
<td>-130.87</td>
<td>13.91</td>
</tr>
<tr>
<td>5</td>
<td>SBCOR</td>
<td>-75.00</td>
<td>1.49</td>
</tr>
<tr>
<td>6</td>
<td>1° derivative FF + LSS</td>
<td>12.50</td>
<td>39.14</td>
</tr>
<tr>
<td>7</td>
<td>1° derivative RSD + LSS</td>
<td>25.00</td>
<td>41.89</td>
</tr>
</tbody>
</table>

Table 1: Relative average improvements of all the reported techniques with respect to the performance of the standard MFCC.