Recognizing Cochlear Implant-like Spectrally Reduced Speech with HMM-based ASR: Experiments with MFCCs and PLP Coefficients

Cong-Thanh Do, Dominique Pastor, Gaël Le Lan and André Goalic

Institut Télécom, Télécom Bretagne, UMR CNRS 3192 Lab-STICC
Technopôle Brest-Iroise, CS 83818, 29238 Brest Cedex 3, France
{thanh.do, dominique.pastor, gael.lelan, andre.goalic}@telecom-bretagne.eu

Abstract

In this paper, we investigate the recognition of cochlear implant-like spectrally reduced speech (SRS) using conventional speech features (MFCCs and PLP coefficients) and HMM-based ASR. The SRS was synthesized from subband temporal envelopes extracted from original clean speech for testing, whereas the acoustic models were trained on a different set of original clean speech signals of the same speech database. It was shown that changing the bandwidth of the subband temporal envelopes had no significant effect on the ASR word accuracy. In addition, increasing the number of frequency subbands of the SRS from 4 to 16 improved significantly the system performance. Furthermore, the ASR word accuracy attained with the original clean speech, by using both MFCC-based and PLP-based speech features, can be achieved by using the 16-, 24-, or 32-subband SRS. The experiments were carried out by using the TI-digits speech database and the HTK speech recognition toolkit.

Index Terms: Spectrally reduced speech, subband temporal envelope, cochlear implant, HMM-based ASR, MFCCs, PLP coefficients.

1. Introduction

It is important to reduce speech signal variability, due to speech production (accent and dialect, speaking style, etc.) or environment (additive noise, microphone frequency response, etc.), in order to guarantee stable ASR performance. Therefore, in an ASR system, the speech analysis module is aimed at reducing the speech signal variability and extracting the ASR relevant spectral information into speech acoustic features. However, despite the speech variability reduction achieved by such standard speech signal analyses, ASR performance is still adversely affected by noise and other sources of acoustic variability [1]. Since most standard speech processing analyses for ASR are performed in the spectral domain, it is natural to seek the relevant spectral information that is sufficient for ASR, in the speech signal.

In ASR based on Hidden Markov Models (HMMs), one way to estimate the ASR relevant spectral information is to evaluate the ASR performance on spectrally reduced speech (SRS) signals when the acoustic models (the HMMs) are trained on a clean speech (full spectrum) database. As usual, the tested signals must not belong to the training database. Such an approach was first investigated by Barker and Cooke in [2] where the authors “consider how acoustic models trained on original clean speech can be adapted to cope with a particular form of spectral distortion, namely reduction of clean speech to sinewave replica[s]”. The ASR results were, however, not satisfactory in train-test unmatched conditions. The cepstral coding techniques are, following the authors, “inappropriate for dealing with drastic alterations to the shape of the spectral profile caused by spectral reduction” [2].

The acoustic simulation of cochlear implant is a spectrally reduced transform of original speech [3]. This type of SRS should be appropriate to evaluate the relevant spectral information needed by an HMM-based ASR whose acoustic models are trained on a given clean speech database and which uses the Mel frequency cepstral coefficients (MFCCs) [4] or the perceptual linear prediction (PLP) coefficients [5] as acoustic features. The rationale is twofold. On the one hand, cochlear implant-like SRS can be recognized by normal hearing listeners. The recognition scores then depend on the spectral resolution (or the number of frequency subbands) of the SRS [3]. Furthermore, human cochlear implant listeners relying on primarily temporal cues can achieve a high level of speech recognition in quiet environment [6]. The foregoing facts suggest that the cochlear implant-like SRS could contain sufficient information for human speech recognition, even though such an SRS is synthesized from the speech temporal envelopes only. On the other hand, in ASR, certain speech analyses, such as the Bark or Mel-scale warping of the frequency axis or the spectral amplitude compression, performed on the conventional speech acoustic features (MFCCs or PLP coefficients), derive from the model of the human auditory system. Such auditory-like analyses, which mimic the speech processing performed by the human auditory system, are basically aimed at reducing speech signal variability and emphasizing the most relevant spectral information for ASR [7]. As a result, the cochlear implant-like SRS should contain sufficient spectral information for ASR based on conventional acoustic features, such as the MFCCs or the PLP coefficients.

In [8], we have shown that the SRS with high spectral resolution contains sufficient spectral information for HMM-based ASR using the MFCCs as speech feature. The experiments were carried out by using the TI-digits speech database [9] and the HTK speech recognition toolkit [10]. In present paper, we summarize the results obtained in [8] and, further, present the ASR results of the SRS with another type of speech feature, the PLP coefficients. In addition, we perform a spectral distortion analysis on the synthesized SRS signals in order to clarify the spectrally reduced effect in this type of speech. The paper is organized as follows. Section 2 presents the SRS synthesis algorithm. Subsequently, the spectral distortion analysis is presented in section 3. Next, the ASR results with MFCC-based and PLP-based speech features are presented in section 4. Finally, the conclusions and perspectives about the automatic recognition of the SRS by using MFCCs and PLP coefficients with HMM-based ASR are proposed in section 5.
2. SRS Synthesis Algorithm

A speech signal is first decomposed into \( N \) subband signals by using a filterbank composed of \( N \) bandpass filters with \( N \) taking values in \( \{4, 8, 16, 24, 32\} \). The analysis filterbank is aimed at simulating the motion of the basilar membrane [11]. In this respect, the filterbank consists of nonuniform bandwidth bandpass filters that are linearly spaced on the Bark scale. In the literature, gammatone filters [12] or elliptic filters [3], have been used to design such a filterbank. In this paper, each bandpass filter in the filterbank is a second-order elliptic bandpass filter having a minimum stop-band attenuation of 50-dB and a 2-dB of peak-to-peak ripple in the pass-band. The lower, upper and central frequencies of the bandpass filters are calculated as in [13]. An example of analysis filterbank is given in Fig. 1.

![Frequency response of an analysis filterbank consisting of 16 second-order elliptic bandpass filters used for speech signal decomposition. The speech signal is sampled at 8 kHz.](image)

The amplitude modulations (AMs) of the subband signals are then extracted by, first, full-wave rectification of the outputs of the bandpass filters and, subsequently, lowpass filtering of the resulting signals. The sampling rate of the AM is kept at the same value as that of the subband signal (8 kHz). In this study, the AM filter is a fourth-order elliptic lowpass filter with 2-dB of peak-to-peak ripple and a minimum stop-band attenuation of 50-dB. The cutoff frequency of the AM lowpass filter is chosen in \( \{16, 50, 160, 500\} \) Hz to evaluate the effect of reducing the bandwidth of the temporal envelope information [3]. The subband amplitude modulation is then used to modulate a sinusoid whose frequency equals the central frequency of the corresponding analysis bandpass filter of that subband. Afterwards, the subband modulated signal is spectrally limited (i.e. is filtered again) by the same bandpass filter used for the original analysis subband [3]. Finally, all the subband spectrally limited signals are summed to synthesize the SRS.

We selected 250 utterances from the clean speech testing database of TI-digits. These 250 utterances were selected so that they were spoken by adult and child speakers of both sexes. The lengths of the utterances in the set vary from the minimum length (isolated digit sequence) to the maximum length (seven-digit sequence) of the sequences in the TI-digits. The SRSs were synthesized from these 250 utterances by using the SRS synthesis algorithm described above. With 5 values \( \{4, 8, 16, 24, 32\} \) of the number of frequency subbands and 4 values \( \{16, 50, 160, 500\} \) Hz for the bandwidth of the subband temporal envelope (AM), we have thus 20 sets of synthesized SRS signals, each set contains 250 SRS utterances. These sets will be used for the spectral distortion analyses and the ASR tests in the next sections.

3. Spectral Distortion Analysis

Given an original clean speech utterance \( x_i \) and the corresponding SRS utterance \( \hat{x}_i \), which is synthesized from \( x_i \), assume that \( f_j \) and \( f_j \) are the spectra of two speech frames of \( x_i \) and \( \hat{x}_i \), respectively. The spectral distortion between these two speech frames can be measured on their spectra \( d_{x_i, \hat{x}_i}(f_j, \hat{f}_j) \). A good spectral distortion measure should have the following properties [14]:

1. \( d_{x_i, \hat{x}_i}(f_j, \hat{f}_j) \geq 0 \), with equality when \( f_j = \hat{f}_j \);
2. \( d_{x_i, \hat{x}_i}(f_j, \hat{f}_j) \) should have a reasonable perceptual interpretation;
3. \( d_{x_i, \hat{x}_i}(f_j, \hat{f}_j) \) should be numerically tractable and easy to compute.

Amongst the available spectral distortion measures [14], the weighted slope metric (WSM) distortion measure [15], which is a perceptually based distortion measure, satisfies well these three properties. This spectral distortion measure reflects the spectral slope difference near spectral peaks in a critical-band spectral representation [14]. It is also shown that the WSM distortion measure correlates well with the perceptual data [15]. Further, the WSM distortion measure is one of the spectral distortion measures that gave the highest recognition score with a standard dynamic time warping (DTW) based, isolated word, speech recognizer [14]. In this respect, we use the WSM distortion measure to assess the spectral distortion between two speech frames extracted from an original clean speech utterance and a synthesized SRS utterance, respectively. The mathematical formula of the WSM distortion measure can be found in [15] and [14]. Henceforth, \( d_{x_i, \hat{x}_i}(f_j, \hat{f}_j) \) designates the WSM distortion measure. The length of the speech frames for the WSM distortion measure is 10 ms. The Matlab programs for calculating the WSM spectral distortion can be found at [16].

The spectral distortion \( \eta_{x_i, \hat{x}_i} \) between two speech utterances \( x_i \) and \( \hat{x}_i \) can be defined as the mean of the WSM spectral distortion measures between the speech frames \( d_{x_i, \hat{x}_i}(f_j, \hat{f}_j) \):

\[
\eta_{x_i, \hat{x}_i} = \frac{1}{M} \sum_{j=1}^{M} d_{x_i, \hat{x}_i}(f_j, \hat{f}_j)
\]

(1)

where \( M \) is the number of speech frames in the speech utterances \( x_i \) and \( \hat{x}_i \). We define the overall spectral distortion \( \bar{\eta} \) as the average of the \( \eta_{x_i, \hat{x}_i} \) calculated for all the \( N = 250 \) utterances in each set of SRS signals:

\[
\bar{\eta} = \frac{1}{N} \sum_{i=1}^{N} \eta_{x_i, \hat{x}_i} = \frac{1}{N} \frac{1}{M} \sum_{i=1}^{N} \sum_{j=1}^{M} d_{x_i, \hat{x}_i}(f_j, \hat{f}_j)
\]

(2)

We then calculate the overall spectral distortion \( \bar{\eta} \) for all 20 sets of SRS utterances (see section 2). The values of \( \bar{\eta} \) are illustrated in Fig. 2. The error bars in Fig. 2 represent the standard deviations of \( \eta_{x_i, \hat{x}_i} \). An ANOVA revealed no significant difference between the overall spectral distortion \( \bar{\eta} \) when the bandwidth of the subband temporal envelopes are changed from 16 Hz to 500 Hz \( [F(3,16) = 0.64, p > 0.5] \). Generally, the overall spectral distortion \( \bar{\eta} \) decreases when the number of
frequency subbands of the SRS increases. Since the WSM distortion measure is perceptually based and is correlated with the perceptual data [15], we can deduce that the value of \( \eta \) reflects more or less the level of perceptual distortion that the listeners have to deal with when listening to the SRS signals. In addition, we can remark that \( \eta_{i,\beta} \) varies intensively when the number of frequency subbands of the SRS is small (4- and 8-subband SRS), since the standard deviations of \( \eta_{i,\beta} \) are large at low spectral resolutions of the SRS (4 and 8 subbands).

4. Automatic Speech Recognition Results

We used the HTK speech recognition toolkit [10] to train a speaker-independent HMM-based ASR system on the TI-digits speech database [9]. TI-digits is a large speech database of more than 25 thousand connected digits sequences spoken by over 300 men, women, and children. This speech database is widely used in the literature to assess ASR algorithms on small-vocabulary tasks [17, 18, 19]. The data were collected in a quiet environment and digitized at 20 kHz. In this study, the data were downsamplwed to 8 kHz. The ASR system used a bigram language model and the acoustic models were the context-dependent three-state left-to-right triphone HMMs. These models were trained by using both MFCCs and PLP coefficients. The output observation distributions were modelled by Gaussian mixture models (GMMs) [20]. In each state, the number of mixture components was 16. The feature vectors consist of 13 MFCCs or 13 PLP coefficients. For the MFCCs and the PLP coefficients calculation, the standard filterbank consisting of 26 filters was used [10]. The MFCCs and the PLPs coefficients were calculated from every Hamming windowed speech frame of 25 ms length and with an overlap of 15 ms between two adjacent frames. The first (delta) and second (acceleration) difference coefficients were appended to the static MFCCs and PLP coefficients to provide 39-dimensional feature vectors. This configuration for the feature vectors was used in both training and testing conditions. Next, the ASR tests were taken on the 20 sets of synthesized SRS signals. The recognition results with the MFCCs and PLP coefficients are shown in Table 1 and Table 2, respectively.

Table 1: ASR word accuracies (in %) computed on the 250 SRSs synthesized from 5 values for the number of frequency subbands and 4 values for the bandwidth of the subband temporal envelopes (AM). The ASR system was trained on the TI-digits clean speech training database. The speech feature vectors were MFCC-based. The ASR word accuracy computed on the 250 original clean speech utterances is 99.76%.

<table>
<thead>
<tr>
<th>AM bandwidth</th>
<th>Number of frequency subbands</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td>16 Hz</td>
<td>94.3</td>
</tr>
<tr>
<td>50 Hz</td>
<td>93.5</td>
</tr>
<tr>
<td>100 Hz</td>
<td>92.7</td>
</tr>
<tr>
<td>500 Hz</td>
<td>91.9</td>
</tr>
</tbody>
</table>

For the ASR results with MFCC-based speech feature vectors, an ANOVA performed on the lines of Table 1 revealed that changing the bandwidth of the subband temporal envelopes had no significant effect in terms of ASR word accuracy [F(3,16) = 0.11, \( p > 0.05 \)]. However, another ANOVA performed on the columns of Table 1 indicated that changing the number of frequency subbands had a significant effect across all tests [F(4,15) = 73.9, \( p < 0.001 \)]. Protected \( t \)-tests (or Least Significant Difference (LSD) tests) [21] were thus performed and showed that the 8-subband SRS yielded significant better ASR word accuracies compared to the 4-subband SRS [\( t_{\text{obs}}(4,8) = 11.23 > t_{\text{crit}}(15) = 4.07, \alpha = 0.001 \)], where \( t_{\text{obs}}(4,8) \) is the protected \( t \)-test value calculated between the 4-subband SRS and the 8-subband SRS word accuracies, \( t_{\text{crit}}(15) \) is the critical value at the desired \( \alpha \) level for 15 degrees of freedom. In contrast, no significant difference was revealed amongst the 16, 24, and 32-subband SRSs in terms of ASR word accuracy [\( t_{\text{obs}}(16,24) \approx t_{\text{obs}}(24,32) \approx t_{\text{obs}}(16,32) \approx 0 < t_{\text{crit}}(15) = 2.13, \alpha = 0.05 \)]. The ASR word accuracies of each SRS in this group are significant better than those of the 8-subband SRS [\( t_{\text{obs}}(8,16) = 2.79, t_{\text{obs}}(8,24) = 2.80, t_{\text{obs}}(8,32) = 2.78, t_{\text{crit}}(15) = 2.13, \alpha = 0.05 \)].

For the ASR results with PLP-based speech feature vectors, the same statistical analyses had been performed on the ASR word accuracies in Table 2 and the same conclusions would be revealed. In summary, increasing the number of the SRS frequency subbands from 4 to 16 made significant improvement in terms of ASR word accuracy. Interestingly, the 16, 24, and 32-subband SRSs could achieve an ASR word accuracy comparable to that attained with the original clean speech.

Table 2: ASR word accuracies (in %) computed on the 250 SRSs synthesized from 5 values for the number of frequency subbands and 4 values for the bandwidth of the subband temporal envelopes (AM). The ASR system was trained on the TI-digits clean speech training database. The speech feature vectors were PLP-based. The ASR word accuracy computed on the 250 original clean speech utterances is 99.70%.

<table>
<thead>
<tr>
<th>AM bandwidth</th>
<th>Number of frequency subbands</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td>16 Hz</td>
<td>94.3</td>
</tr>
<tr>
<td>50 Hz</td>
<td>93.5</td>
</tr>
<tr>
<td>100 Hz</td>
<td>92.7</td>
</tr>
<tr>
<td>500 Hz</td>
<td>91.9</td>
</tr>
</tbody>
</table>

For the ASR results with PLP-based speech feature vectors, an ANOVA performed on the lines of Table 1 revealed that changing the bandwidth of the subband temporal envelopes had no significant effect in terms of ASR word accuracy [F(3,16) = 0.11, \( p > 0.05 \)]. However, another ANOVA performed on the columns of Table 1 indicated that changing the number of frequency subbands had a significant effect across all tests [F(4,15) = 73.9, \( p < 0.001 \)]. Protected \( t \)-tests (or Least Significant Difference (LSD) tests) [21] were thus performed and showed that the 8-subband SRS yielded significant better ASR word accuracies compared to the 4-subband SRS [\( t_{\text{obs}}(4,8) = 11.23 > t_{\text{crit}}(15) = 4.07, \alpha = 0.001 \)], where \( t_{\text{obs}}(4,8) \) is the protected \( t \)-test value calculated between the 4-subband SRS and the 8-subband SRS word accuracies, \( t_{\text{crit}}(15) \) is the critical value at the desired \( \alpha \) level for 15 degrees of freedom. In contrast, no significant difference was revealed amongst the 16, 24, and 32-subband SRSs in terms of ASR word accuracy [\( t_{\text{obs}}(16,24) \approx t_{\text{obs}}(24,32) \approx t_{\text{obs}}(16,32) \approx 0 < t_{\text{crit}}(15) = 2.13, \alpha = 0.05 \)]. The ASR word accuracies of each SRS in this group are significant better than those of the 8-subband SRS [\( t_{\text{obs}}(8,16) = 2.79, t_{\text{obs}}(8,24) = 2.80, t_{\text{obs}}(8,32) = 2.78, t_{\text{crit}}(15) = 2.13, \alpha = 0.05 \)].

For the ASR results with PLP-based speech feature vectors, the same statistical analyses had been performed on the ASR word accuracies in Table 2 and the same conclusions would be revealed. In summary, increasing the number of the SRS frequency subbands from 4 to 16 made significant improvement in terms of ASR word accuracy. Interestingly, the 16, 24, and 32-subband SRSs could achieve an ASR word accuracy comparable to that attained with the original clean speech.
signals (the ASR word accuracies computed on the 250 original clean speech utterances with MFCC-based and PLP-based speech feature vectors were 99.76% and 99.70%, respectively).

5. Conclusion and Perspective

In this paper, we have investigated the automatic recognition of cochlear implant-like SRS, which is synthesized from subband temporal envelopes of the original clean speech. The MFCCs and the PLP coefficients were used as speech features and the ASR system, which is speaker-independent and HMM-based, was trained on the original clean speech training database of TI-digits. It was shown that changing the bandwidth of the subband temporal envelopes had no significant effect on the ASR system word accuracy. In addition, increasing the number of frequency subbands of the SRS from 4 to 16 improved significantly the performance of the ASR system. However, there was no significant difference amongst the 16, 24, and 32-subband SRS in terms of ASR word accuracy. It was possible to achieve an ASR word accuracy as good as that attained with the original clean speech by using SRS with 16, 24, or 32 frequency subbands and by using both MFCC-based and PLP-based speech features.

The MFCCs or the PLP coefficients, along with the delta and acceleration coefficients, were concatenated together in the acoustic feature vector and were used in both the training and the testing conditions. The results presented in this paper suggest that SRS with 16, 24, and 32 subbands contain sufficient spectral information for speech recognition with HMM-based ASR system using the MFCCs or the PLP coefficients along with the delta and acceleration coefficients. The SRS and original clean speech are quite different in the signal domain since the SRS is synthesized from original clean speech temporal envelopes, only. The spectral distortion analysis, based on the WSM distortion measure, showed that there are significant spectral distortions in the synthesized SRS compared to the original clean speech signal. Despite of these spectral distortions which are induced by the SRS synthesis, and although the ASR acoustic models are trained on the clean speech training database (TI-digits), the ASR word accuracy, computed on original clean speech signals of a testing set of TI-digits, is still maintained with the synthesized SRS of 16, 24, and 32 subbands.

Therefore, the 16-, 24-, and 32-subband SRS models might lead to the design of new acoustic features and suggest new speech models, for ASR, that could be robust to noise and other sources of acoustic variability. In this respect, performing ASR with HMMs trained on SRS could assess the relevance of SRS along with the delta and acceleration coefficients. The SRS and the testing conditions. The results presented in this paper strongly suggest that SRS with 16, 24, and 32 subbands contain sufficient spectral information for speech recognition with HMM-based ASR system using the MFCCs or the PLP coefficients along with the delta and acceleration coefficients. The SRS and original clean speech are quite different in the signal domain since the SRS is synthesized from original clean speech temporal envelopes, only. The spectral distortion analysis, based on the WSM distortion measure, showed that there are significant spectral distortions in the synthesized SRS compared to the original clean speech signal. Despite of these spectral distortions which are induced by the SRS synthesis, and although the ASR acoustic models are trained on the clean speech training database (TI-digits), the ASR word accuracy, computed on original clean speech signals of a testing set of TI-digits, is still maintained with the synthesized SRS of 16, 24, and 32 subbands.

Therefore, the 16-, 24-, and 32-subband SRS models might lead to the design of new acoustic features and suggest new speech models, for ASR, that could be robust to noise and other sources of acoustic variability. In this respect, performing ASR with HMMs trained on SRS could assess the relevance of SRS as a model for speech recognition. In addition, the fact that cochlear implant-like SRS is synthesized from speech temporal envelopes, only, might help in reducing ASR irrelevant spectral information due to environment. In future work, we plan to consider noisy speech in which the spectral information is degraded by noise. We expect that the detrimental effect of the degraded speech spectral information on ASR can be partially reduced by means of the SRS synthesis used in this paper.

6. References