FEATURE EXTRACTION FOR ROBUST AUTOMATIC SPEECH RECOGNITION USING SYNCHRONY OF ZERO CROSSINGS

Richard M. Stern, Pierre Ponce, and Rita Singh
Department of Electrical and Computer Engineering and School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213 USA

ABSTRACT

This paper describes a new method of extracting features for automatic speech recognition systems that are based on the temporal synchrony of zero crossings of the speech waveform. This feature extraction procedure, referred to as the ZSYNC method, is motivated by the inherent robustness of zero-crossing information to additive noise, and by the greater observed robustness of synchrony-based physiological responses to speechlike sounds compared to representations based on mean rate of response. ZSYNC feature vectors are obtained by passing the speech signal through a bank of bandpass filters, determining the locations of positive zero crossings of the filter outputs, and computing the normalized synchrony measure of these zero crossings. Feature vectors similar to cepstral coefficients are obtained by computing the discrete cosine transform of the log of the product of synchrony and short-term power. In preliminary experiments using several standard speech databases, speech recognition accuracy using these features exhibits a modest but very consistent improvement over the recognition accuracy obtained with conventional Mel-frequency cepstral coefficients.

1. INTRODUCTION

Most current automatic speech recognition systems use cepstrum-based features such as mel-frequency cepstral coefficients (MFCC) [1] or perceptual linear prediction (PLP) [4] cepstral coefficients, that provide an estimate of short-term energy as a function of frequency. In these cases and their many variations, features are derived based only on time \textit{averages} of the waveform after appropriate filtering and transformation and the detailed \textit{time structure} is lost. In this paper we describe a new pragmatic feature extraction approach that is based on the synchrony of temporal fine structure to the analysis frequencies of the filterbank analysis. We refer to this type of processing as “zero-crossing synchrony”, or ZSYNC processing, and the resulting features as ZSYNC features.

Speech recognition experiments conducted on several databases recorded under different acoustic conditions indicate that the ZSYNC features result in a modest, but consistent improvement in recognition performance over conventional MFCC and PLP features. Experimental results also indicate that further gains in recognition performance can be obtained by combining the hypotheses obtained from ZSYNC and MFCC features.

In the following section we describe the various motivations behind ZSYNC processing. In Sec. 3 we describe in some detail how the ZSYNC features are extracted, and in Sec. 4 we compare speech recognition results using the ZSYNC features with conventional MFCC and PLP features. Finally, in Sec. 5 we present our conclusions.

2. MOTIVATION FOR ZERO-CROSSING SYNCHRONY PROCESSING

There are several motivations for reconsidering the use of temporal features, and in particular zero crossings as the basis of feature extraction. Intuitively, zero crossings are appealing as the basis for features because the values of zero crossings are far less affected by additive noise than amplitude-based features because of the slope of the waveform tends to be greatest near the zero-crossing point. As an example, De Mori [2] has demonstrated both mathematically and empirically that the use of zero-crossing information can provide more robust estimation of features based on spectral center of gravity.

Our use of the synchrony of zero crossings has also been motivated by the results of physiological studies of the representation of speech sounds at the level of the auditory nerve. For example, Sachs and Young [7, 11] have compared the representations of synthetic vowel-like stimuli using the mean rate of response to a synchrony-based measure, the “average localized rate” (ASLR). They found that the measured ASLR in the response of the auditory-nerve fibers to these stimuli was far less affected by additive noise than their mean rate of response.

The work of Sachs and Young was one of the inspirations for the synchrony measurement that is a part of many computational models of auditory processing such as Seneff’s Generalized Synchrony Detector (GSD) [8]. Seneff has demonstrated that the use of synchrony-based information provides a much clearer visual display of speech information, and representations such as hers have been used as the basis for feature extraction in several speech recognition systems.

3. DESCRIPTION OF ZERO-CROSSING SYNCHRONY PROCESSING

The ZSYNC features can be thought of as a computational implementation of cepstral features derived from an estimate of Sachs and Young’s ASLR measure as derived from waveform zero crossings rather than auditory-nerve spike trains.

The speech waveform is multiplied by a series of Hamming windows of duration 25 ms in the usual fashion. In each frame the windowed waveform is passed through a bank of 40 Gammatone bandpass filters [6] which were implemented as in Slaney’s Auditory Toolbox. The filters had center frequencies spaced according to the equivalent rectangular bandwidth (ERB) scale. The positive zero crossings of the filter outputs were recorded.
An estimate of the local synchrony was obtained by tabulating histograms of the zero crossing intervals relative to a time interval equal to the reciprocal of the center frequency of each bandpass channel. Synchrony is then estimated by computing the ratio of the first Fourier component of the phase histogram to the zero-th Fourier component, following the procedure of Sachs and Young. This computation produces a number between 0 and 1 which represents the extent to which the output is synchronized to the bandpass channel frequency. This normalized synchrony coefficient is then multiplied by the short-term energy of the signal emerging from the bandpass filter channel in the given frame.

This product of filter energy by synchrony coefficient replaces the outputs of the Mel filters in conventional MFCC processing. Further computation follows that of conventional mel frequency cepstral coefficients, in that the discrete cosine transform is performed on the logarithms of the energy-normalized synchrony measures. A small threshold value is added to the synchrony coefficient before the log operation to reduce the impact of frames when the synchrony is exactly zero on the features obtained.

The feature extraction used in this processing is similar in some ways to Ghitza’s ensemble interval histogram (EIH) model [3], although the subsequent processing is different.

4. EXPERIMENTAL RESULTS

We describe in this section comparisons of the recognition accuracy obtained using the ZSYNC features to those of conventional MFCC and PLP processing. It should be borne in mind that these results are preliminary, as we have not yet had the opportunity to optimize any of the parameters associated with the ZSYNC feature extraction.

4.1. DARPA Resource Management Database

Experiments were run using the DARPA Resource Management (RM) database [Price 1988] on the CMU Sphinx-III HMM-based speech recognition system [9]. Acoustic models were trained with MFCC and ZSYNC features derived from 2880 utterances of uncorrupted speech. The test set consisted of 1600 utterances from the RM database. For the context-independent (CI) models each HMM state was modeled by a single Gaussian density. Context-dependent (CD) triphone models were also trained using 2000 tied states, where each state was represented by a mixture of 4 Gaussians. For both MFCC-based and ZSYNC-based systems, a combination of 13 cepstra, 13 delta cepstra, and 13 double-delta cepstra for a total of 39 features was used. Since the linguistic structure of the RM database is highly constrained, a lower-than-optimal language weight was used during recognition, so that the contributions of the acoustic models and choice of feature set were better reflected in the results obtained.

Recognition accuracy obtained using the RM database is summarized in Table 1. We note that the use of the ZSYNC features provides an improvement of 4.3% and 5.9% for the CI and CD models respectively.

4.2. NRL SPINE Database

The first “Speech in Noisy Environments” (SPINE1) evaluation was conducted by the Naval Research Laboratories (NRL) in August, 2000. The purpose of the evaluation was to provide impetus to the design of algorithms and strategies which improve the performance of speech recognition systems in the presence of varied noises. The task consisted of recognizing approximately nine hours of speech from battleship games with realistic military noises playing in the background. Approximately eight hours of speech recorded under similar conditions were provided for training the systems. The training data had fewer speakers and noise types than the evaluation data. The signal-to-noise ratio (SNR) for both the training and test data varied from 5 dB to 20 dB.

CMU implemented several systems for the SPINE task, as described in [10]. After the SPINE evaluation was completed we ran a subsequent series of comparisons of the performance of MFCC, PLP, and ZSYNC features through only the early stages of the CMU SPINE primary system.

<table>
<thead>
<tr>
<th>Feature Set WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC 35.1</td>
</tr>
<tr>
<td>PLP 38.0</td>
</tr>
<tr>
<td>ZSYNC 34.9</td>
</tr>
<tr>
<td>MFCC+PLP 32.8</td>
</tr>
<tr>
<td>MFCC+ZSYNC 32.6</td>
</tr>
<tr>
<td>PLP+ZSYNC 33.4</td>
</tr>
<tr>
<td>MFCC+PLP+ZSYNC 31.7</td>
</tr>
</tbody>
</table>

Table 2: Comparison of WER obtained for the NRL SPINE task using various combinations of MFCC, PLP, and ZSYNC features.

Table 2 compares the recognition accuracy obtained using various combinations of the three feature sets. Features were combined at the hypothesis level by collapsing recognition hypotheses obtained separately from all features into a word graph, which was then rescored, as described in [10]. We note that the use of the ZSYNC features produced relative reductions in WER of 0.6% and 8.1% compared to MFCC and PLP features, respectively. The use of ZSYNC features improved recognition accuracy by 7.1% when used in combination with MFCC features alone, and by 9.7% when used in combination with MFCC and PLP features together. We believe that these further improvements in recognition accuracy suggest that the ZSYNC features provide information about the speech waveform that is complementary to that provided by the other feature sets.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>CI Models</th>
<th>CD Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>23.8</td>
<td>10.2</td>
</tr>
<tr>
<td>ZSYNC</td>
<td>22.2</td>
<td>9.6</td>
</tr>
</tbody>
</table>

Table 1: Comparison of word error rates (WER) obtained using MFCC and ZSYNC cepstral features on the DARPA Resource Management Database.
4.3. Telefónica Database

A third set of experiments was performed using a proprietary database developed by Telefónica Investigación y Desarrollo (TID). This corpus was recorded over the GSM cellular telephone network in Spain and represents an important effort in the research of cellular telephony speech recognition. It is a small-vocabulary corpus of numbers and quantities (with a lexicon of approximately 75 words). There were approximately 6000 training utterances and 3000 testing utterances (with 13091 tokens in the test set), spoken by approximately 1700 speakers. It was originally recorded at 8000 kHz. The database has been manually transcribed and annotated according to acoustic environment, speaker dialect and other conditions. It contains a broad sample of speaker and environmental conditions.

We used a modified version of SPHINX-3 that had been developed originally for Spanish-language broadcast news transcription [5] to compare the performance of the various feature sets discussed in this paper on the TID database. Semi-continuous 3-state HMMs were trained with MFCC, PLP, and ZSYNC features. In all cases context-dependent triphone models with 600 tied states were used.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>9.0</td>
</tr>
<tr>
<td>PLP</td>
<td>10.0</td>
</tr>
<tr>
<td>ZSYNC</td>
<td>8.8</td>
</tr>
<tr>
<td>MFCC+ZSYNC</td>
<td>8.5</td>
</tr>
<tr>
<td>MFCC+PLP+ZSYNC</td>
<td>8.7</td>
</tr>
</tbody>
</table>

Table 3: Comparison of WER for the Telefónica task using combinations of the MFCC, PLP, and ZSYNC features.

Table 3 summarizes the WER obtained using the TID database and various combinations of the MFCC, PLP, and ZSYNC features. We note that the ZSYNC features provided a relative reduction in WER of 2.2% and 12% compared to the MFCC and PLP features, respectively. Combining the ZSYNC hypotheses with MFCC hypotheses provides an additional reduction in WER of 5.6% relative to MFCC alone.

5. SUMMARY AND CONCLUSIONS

In this paper we have described a new set of features for use in automatic speech recognition that are based on the synchrony of zero crossings of bandpass-filtered speech data to the center frequencies of the filters. The data considered represent a wide variety of environmental and phonetic conditions. In all cases the ZSYNC features provided better recognition accuracy than the conventional MFCC and PLP features, although the differences were small in some cases. Further and more substantial improvements were observed when the ZSYNC features were combined with other feature sets. It should be noted that all of the results described are preliminary and were obtained without any parameter optimization. We strongly anticipate that performance will continue to improve with further study and analysis of the ZSYNC features and their properties.

ACKNOWLEDGEMENTS

This research was sponsored by the Department of the Navy, Naval Research Laboratory under Grant No. N00014-93-1-2005. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

REFERENCES

A note from the authors to the reviewers: This paper was written under great time pressure, which might be evident from the shortness of the manuscript. While we expect the paper to be reviewed on the basis of what is contained in this report, please be assured that we will submit a revision before January 22 with far more detailed description and explanation and illustrative examples, along with far more thoughtful analyses.