Synchrony-Based Feature Extraction for Robust Automatic Speech Recognition

Fernando de-la-Calle-Silos and Richard M. Stern, Fellow, IEEE

Abstract—This paper discusses the application of models of temporal patterns of auditory-nerve firings to enhance robustness of automatic speech recognition systems. Most conventional feature extraction schemes (such as MFCC and PLP coefficients) are based on short-time energy in each frequency band, and the temporal patterns of auditory-nerve activity are discarded. We compare the impact on speech recognition accuracy of several types of feature extraction schemes based on the putative synchrony of auditory-nerve activity, including feature extraction based on a modified version of the generalized synchrony detector (GSD) proposed by Seneff, and a modified version of the averaged localized synchrony response (ALSR) proposed by Young and Sachs. It was found that the use of features based on auditory-nerve synchrony can indeed improve speech recognition accuracy in the presence of additive noise based on experiments using multiple standard speech databases. Recognition accuracy obtained using the synchrony-based features is further increased if some form of noise removal is applied to the signal before the synchrony measure is estimated. Signal processing for noise removal based on the noise suppression that is a part of PNCC feature extraction is more effective toward this end than conventional spectral subtraction.

Index Terms—Robust speech recognition, feature extraction, physiological modeling, auditory modeling, auditory synchrony.

I. INTRODUCTION

Performance in automatic speech recognition (ASR) tasks is still far worse than that of human speech recognition, and noisy or reverberant environments only compound the problem. Many researchers over the years have suggested that feature extraction techniques motivated by processing in the human auditory system may be useful in reducing this gap in performance (e.g. [1], [2]). For example, the commonly-used Mel-Frequency Cepstral Coefficients (MFCC) [3] and Perceptual Linear Processing (PLP) features [4], as well as Gammatone-based Coefficients (GTC) [5] and Power-Normalized Cepstral Coefficients [6], [7], result from nonlinear transformations of the frequency domain, and they include a filterbank that mimics the putative bandpass processing in the cochlea. Some other attributes, such as the nonlinear function that relates physical intensity to perceived loudness perception of sound intensity, are also included in MFCC, PLP, and GTC features. In this paper we discuss various ways of exploiting the temporal patterns of auditory-nerve activity to improve ASR accuracy.

Numerous physiological studies have demonstrated that the response of an auditory-nerve fiber with a low characteristic frequency (CF) roughly follows the shape of the input signal at least when the signal amplitude is positive [8]. This “phase-locking” behavior enables the auditory system to compare arrival times of signals to the two ears at low frequencies, which is the basis for the spatial localization of a sound source at these frequencies. While this sort of temporal coding is clearly important for binaural sound localization, it may also play a role in the robust interpretation of signals from individual ears as well. Much of our own work in this area is motivated by physiological findings by Sachs and Young [9] which showed that the average localized synchrony rate (ALSR) that is derived from the nerve firing times is much more robust to changes in intensity of vowel-like sounds than the corresponding mean rate of response as a function of CF. These results suggest that the timing information associated with the response to low-frequency components of a signal can be substantially more robust to variations in intensity (and potentially various other types of signal variability) than the mean rate of the neural response. Most conventional feature extraction schemes (such as MFCC and PLP coefficients) are based on short-time energy in each frequency band, which is associated with mean rate rather than synchrony.

The remainder of this paper is organized as follows: Sec. II briefly reviews the state of the art that has motivated our formulation, Sec. III describes our synchrony measurements and feature extraction procedures in some detail, Sec. IV describes our experimental results, and Sec. V summarizes our findings.

II. BACKGROUND

In this section we very briefly review selected prior studies that describe techniques that have been proposed to develop a “synchrony spectrum” that reflects the temporal patterns of the auditory-nerve response to signals as a function of frequency.

One of the first such descriptions was Seneff’s auditory model [10]. The original formulation included 40 recursive linear filters to mimic auditory-nerve responses (e.g. [11]). Seneff’s model included a four-stage inner hair cell model that described rectification, short-term adaptation, synchrony suppression at higher frequencies, and an automatic gain control (AGC) to normalize response rates. The model had two parallel outputs, one that approximated the instantaneous mean rate of firing and a second that measured the synchrony in response to the incoming signals.

A second early formulation was Ghita’s Ensemble Interval Histogram (EIH) model [12], which develops synchrony information by recording level crossings of previous stage over a set
of seven logarithmically-spaced thresholds over the dynamic range of each channel.

In subsequent years the approaches of Seneff and Ghitza have been elaborated upon, and other techniques have been introduced as well. For example, Ali et al. [13] proposed a simple but useful extension of the Seneff GSD model that develops a synchrony spectrum by simply averaging the responses of several GSDs tuned to the same frequency using inputs from bandpass filters with CFs in a small neighborhood about a central frequency. D.-S. Kim et al. [14] proposed a type of processing called zero-crossing peak analysis (ZCPA), which develops histograms of times between zero crossings weighted by the amplitude of the peak between them. Other recent features motivated by synchrony include SYDOCC features [15] and LNCC features [16], and the signal processing method described by C. Kim et al. [17].

III. SYNCHRONY FEATURE EXTRACTION

In this section we briefly describe the various approaches to synchrony extraction that will be used in the experiments below, focussing on the Generalized Synchrony Detector (GSD) proposed by Seneff [10], along with a second approach that combines the Averaged Localized Synchrony Rate (ALSR) proposed by Young and Sachs [9] with mean rate information. We also discuss the potential benefit that is obtained when synchrony extraction is preceded by a noise cancellation mechanism.

A. Application of the Seneff auditory model and GSD

The Seneff auditory model [10] is well known and has received a great deal of attention in the literature. It contains a model for the auditory-nerve response to sound with two outputs, one representing mean rate and one representing synchrony. Synchrony is estimated via the GSD, which compares the putative instantaneous output of the hair cells in each channel with itself delayed by the reciprocal of the center frequency in each channel; the short-time averages of the sums and differences of these two functions are divided by one another. A threshold is introduced to suppress the response to low-intensity signals and the resulting quotient is passed through a saturating half-wave rectifier to limit the magnitude of the predicted synchrony. In our experience performance is improved by modifying the GSD detector by computing only the inverse of the difference between the original input signal and the original signal delayed by the period of the frequency to which the GSD is tuned. Figure 1 compares the structure of the original and modified GSD calculation; the modified GSD algorithm eliminates the connections denoted by the broken lines, as summarized in the lower portion of Fig. 2.

B. Synchrony estimation based on Averaged Localized Synchronized Rate (ALSR)

We also estimated the synchronized response using a variation of the Averaged Localized Synchronized Rate (ALSR) of

### Footnotes

[17] More specifically, the PNCC-based noise subtraction is accomplished as follows: We retain the original phase and modify only the magnitude spectrum. For each time-frequency bin, following the notation of [2], we obtain the weighting coefficient $w[m,l]$ for the $m$th frame and $l$th frequency band as a ratio of the processed power $T[m,l]$ (the output of the medium-time and short-time PNCC processing) to the original power $P[m,l]$. Each of these channels is associated with $H_l(e^{j\omega_k})$, the frequency response of one of a set of gammatone filters. The final spectral weighting $\mu[m,k]$ is obtained using the above weight $w[m,l]$ according to the equation:

$$\mu[m,k] = \frac{\sum_{l=0}^{L-1} w[m,l] |H_l(e^{j\omega_k})|}{\sum_{l=0}^{L-1} |H_l(e^{j\omega_k})|},$$

$$0 \leq k \leq \frac{N}{2}, 0 \leq l \leq L - 1 \quad (1)$$

[17] We noted in our original experiments that recognition accuracy using the GSD could be improved through the use of a noise cancellation mechanism prior to the extraction of synchrony. We considered two types of noise-cancellation approaches in our work. The first, and simpler, approach was to use a form of conventional spectral subtraction [19], but on a band-by-band basis after the initial gammatone filtering [20], as summarized in the upper portion of Fig. 2.

A second approach to noise removal incorporates the nonlinear Asymmetric Noise Suppression (ANS) components of PNCC coefficients [6], [7]. In brief, the speech signal is passed through most of the steps of PNCC processing in order to remove the noise components, and then the audio signal is recovered using spectral reshaping. The enhanced audio signal is then passed through the Seneff front end with the modified GSD, as summarized in the lower portion of Fig. 2.

More specifically, the PNCC-based noise subtraction is accomplished as follows: We retain the original phase and modify only the magnitude spectrum. For each time-frequency bin, following the notation of [2], we obtain the weighting coefficient $w[m,l]$ for the $m$th frame and $l$th frequency band as a ratio of the processed power $T[m,l]$ (the output of the medium-time and short-time PNCC processing) to the original power $P[m,l]$. Each of these channels is associated with $H_l(e^{j\omega_k})$, the frequency response of one of a set of gammatone filters. The final spectral weighting $\mu[m,k]$ is obtained using the above weight $w[m,l]$ according to the equation:

$$\mu[m,k] = \frac{\sum_{l=0}^{L-1} w[m,l] |H_l(e^{j\omega_k})|}{\sum_{l=0}^{L-1} |H_l(e^{j\omega_k})|},$$

$$0 \leq k \leq \frac{N}{2}, 0 \leq l \leq L - 1 \quad (1)$$
The enhanced speech $\hat{x}[n]$ is re-synthesized from the reconstructed spectrum using the overlap-add (OLA) method [22]. Resulting enhanced speech is processed by the Seneff auditory model and the modified GSD, as described above. The resulting synchrony spectrum is subjected to a final DCT which produces a set of coefficients that are similar to cepstral coefficients.

**IV. EXPERIMENTAL RESULTS**

Three standard speech corpora were used for our evaluations: DARPA Resource Management (RM), Wall Street Journal WSJ0 (WSJ0), and Motorola Aurora 4 databases. Since we are concerned primarily with the relative performance of the various signal processing schemes considered, no attempt was made to fine tune the parameters of the SPHINX trainer and decoder to minimize the absolute error rate. Cepstral coefficients $C_0$ to $C_{12}$ were retained together with their corresponding delta ($\Delta$) and acceleration ($\Delta \Delta$) coefficients to yield feature vectors of 39 components. Mean and variance normalizations were applied by utterance on each of the components.

To test the impact of the different methods of synchrony extraction on robustness in recognition accuracy for the RM and WSJ databases we used the same four standard testing environments as in [6]: (1) white noise, (2) noise recorded live on urban streets, (3) single-speaker interference and (4) background music. The street noise was recorded on streets with steady but moderate traffic. The masking signal used for single-speaker-interference experiments consisted of other utterances drawn from the same database as the target speech, and background music was selected from music segments from the original DARPA Hub 4 Broadcast News database.

The various front ends were tested on versions of the test set to which the previously-mentioned noises were added to the corresponding clean speech at four different SNRs using the FANT tool [23] with G.712 filtering. Most evaluations are performed under mismatched conditions (i.e. training on clean speech and testing on degraded speech).

Aurora 4 [24] is a medium-vocabulary task based on the Wall Street Journal (WSJ0) corpus. The experiments were performed using the 16-kHz clean and multi-condition training sets. The evaluation set is derived from WSJ0 5K test set corrupted by six different noises (street traffic, train station, car, babble, restaurant, airport) at 10-20 dB SNR, creating a total of 14 test sets. Note that the types of noise are common
Figure 6 compares results obtained using the GSD processing for matched and mismatched training using Aurora 4 data, averaging over four noise conditions.

Table I provides selected results obtained using the Aurora 4 database under the conditions described above, reporting averages over all the test sets. We note that the use of modified GSD processing with PNCC-based noise subtraction provides relative improvements in WER compared to PNCC processing by 9.1% in mismatched conditions and 6.2% for matched conditions. Improvements compared to MFCC features are much greater, of course.

V. SUMMARY AND CONCLUSIONS

In this paper we compared the improvements in speech recognition accuracy that can be obtained through the use of several types of features that are based on the extent to which the auditory-nerve representation of a signal is synchronized in its response. The most effective synchrony-based feature was a modified version of the Seneff Generalized Synchrony Detector preceded by noise removal based on PNCC processing. This feature provided substantially better recognition accuracy than baseline PNCC features for speech that is degraded by white noise, interfering speakers, and reverberation. Improvements for speech in the presence of street noise and background music were more modest. We are attempting to develop more efficient ways of combining the noise removal provided by PNCC processing with the synchrony representation of GSD processing.