A Feature Extraction Method for Automatic Speech Recognition Based on the Cochlear Nucleus

Serajul Haque, Roberto Togneri

School of Electrical, Electronic and Computer Engineering, University of Western Australia, {serajul}{roberto}@ee.uwa.edu.au

Abstract
Motivated by the human auditory system, a feature extraction method for automatic speech recognition (ASR) based on the differential processing strategy of the AVCN, PVCN and the DCN of the cochlear nucleus is proposed. The method utilizes a zero-crossing with peak amplitudes (ZCPA) auditory model as synchrony detector to discriminate the low frequency formats. It utilizes the mean rate information in the synapse processing to capture the very rapidly changing dynamic nature of speech. Additionally, a temporal companding method is utilized for spectral enhancement through two-tone suppression. We propose to separate synchrony detection from synaptic processing as observed in the parallel processing methodology in the cochlear nucleus. HMM recognition using isolated digits showed improved recognition rates in clean and in non-stationary noise conditions than the existing auditory model.

Index Terms: Speech recognition, zero-crossings, auditory model, cochlear nucleus, hidden Markov model.

1. Introduction
Despite significant advances in ASR in the past three decades, one basic disadvantage is that the performance degrades rapidly in speaker and channel mismatch conditions, and in noisy environments. On the other hand, the process of human speech perception is very much immune to such adverse conditions. This fact has motivated researchers in recent years to emphasize on “human engineering”, that is, to adopt the processing strategies of the human auditory perception in ASR. The application of such human perceptual features may improve ASR performance which has been established in the literature [1], [2].

The knowledge about the peripheral auditory system have been used in auditory models primarily based on psychoacoustic experiments and physiological measurements on frequency selectivity, loudness perception, short-term adaptation, two-tone suppression, lateral inhibition, masking, and other topics. Although much has been known about the peripheral auditory system, comparatively little is known about the neural mechanisms of the central auditory processing stages. It is believed that the initial process in the peripheral auditory system enhances some perceptual cues in the speech which assists the higher auditory system in the proper identification of the speech segments. In response to the initial process, the auditory brainstem provides multiple representations of such information [3].

The auditory nerve (AN) fibres attached to the organ of corti are connected to the next higher region called the cochlear nucleus (CN), located in the central dorso-lateral side of the brainstem. In the CN, specific regions are assigned to specific tasks related to processing speech stimulus, and its principle cells constitute separate, parallel processing pathways for encoding different properties of the auditory signal. In contrast to the AN whose responses are monotype (“primary” response), there are up to 6 different responses observed in the CN [4]. Some cells are excitatory and some are inhibitory. Some pathways provide immunity to noise. Twenty-two different types of neurons have been distinguished in the CN grouped as primary-like, chopper, onset and octopus, build-up and pauser, etc. [4], [3].

At least three major divisions of the CN can be distinguished on the basis of morphology, as shown in Fig. 1. The anteroventral cochlear nucleus (AVCN), innervated by the ascending pathway, consists of primary-like neurons (bushy cells) which can maintain synchrony (phase-lock property) of the AN fibres, transmitting timing information from the AN to more central areas in the auditory system [4], [3]. The posteroventral region of the CN (PVCN) is composed of onset (octopus cells) and chopper neurons which maintain the rate-place code in the AN fibres. The onset cells respond with great precision to signal onsets and broadband transients. Another cell type found in the PVCN are the stellate cells (chopper cells), with their ability to fire a regularly-spaced train of action potentials in tonal or noise stimulus. The more complex dorsal cochlear nucleus (DCN) consists of fusiform cells and the giant cells (also known as type IV cells). Current auditory models of the DCN employ the two-inhibitor model in which type IV cells receive excitation directly from the auditory nerve, and are inhibited by type II (vertical) cells and a wideband inhibitor (onset-c cells) [6]. The two-tone suppression are reminiscent of the consequences of lateral inhibition used in speech enhancement [7], [5].

In an auditory model used as an ASR front-end, features are extracted from the action potentials generated in the form of a spike train at the AN fiber. However, experiments from auditory physiology demonstrate that higher auditory centres in the brain make use of both rate and synchrony information.

This research is partly funded by the Australian Research Council (ARC) grant No. DP1096348.
Although pitch and loudness are important perceptual attributes, a collection of distinct spectral representations, each of which is adapted to a specific assigned task, is to be preferred over a single complex feature representation which preserves all of the signal attributes [9]. The differential functionality of the higher auditory system has been applied earlier in several cochlear models [9],[10], e.g. Seneff’s joint synchrony/mean rate model [9] provided two separate outputs for the synchrony and the mean rate, each of which represented spectral segments appropriate for distinct subtasks of a speech processing system. However, ASR performance of this model has not been reported. Xiang’s CN model [5] utilizes an IHC-synapse model similar to the stage II of Seneff’s model. The synchrony measure (AVCN processing) is extracted from the power spectrum as weighted average localized synchronized rate (ALSR) cepstra, and mean rate (PVCN processing) as weighted firing rate cepstra. It utilizes a method of weighted ALSR to mimic the lateral inhibition observed in the pauser/build-up neurons (DCN processing). It is applied to speaker identification using modular tree classifiers. It’s performance in ASR is also not reported.

In this paper, a feature extraction method based on the processing strategy of the CN as in Fig. 1 is presented. In an ASR front-end, speech features are extracted from the discharge pattern with all perceptual features integrated into a single set of feature vectors. In contrast, we propose a feature extraction based on the differential processing strategy of the CN. The proposed method has three distinctive features compared to implementations in [9],[10] and [5]. Firstly, the proposed scheme was implemented using temporal processing of speech using a zero-crossing algorithm instead of spectral domain processing. auditory functions are usually dynamic in nature with substantial non-stationary transient properties, which, unlike steady-state emissions, can not be fully characterized by the spectrum. Secondly, a more detailed parallel processing strategy as observed in the CN using multiple filterbanks with wider windows and time constants appropriate for auditory functions were utilized. Thirdly, unlike in [9], synchrony detection was separated from the synaptic processing since synaptic output is more related to dynamic properties than to formant estimates, and that half-wave rectification (HWR) in the synchrony path introduces harmonic distortions which degrades ASR performance. It is also a natural consequence of the differential processing of the CN.

The rest of the paper is arranged as follows. In section 2, we present the CN computational model for ASR. Section 3 introduces the feature extraction by the ZCPA auditory model as a synchrony detector. The lateral inhibition by two-tone suppression is presented in Section 4. The mean rate and the synchrony processes are presented in section 5. Finally, HMM recognition results with isolated Tidigits in stationary and non-stationary additive noise are stated in section 6, followed by conclusions drawn from the experiments in section 7.

2. The CN auditory model

Instead of detailed modeling of the CN, we have attempted a broad functional representation based on the above processing strategies for ASR application. The three different parallel processing algorithms, each one implementing a distinct subtask of the speech recognition system, were implemented. These are for the synchrony detection, the mean discharge rate processing, and the two-tone suppression across the frequency range, as observed in the AVCN, the PVCN and the DCN, respectively. The primarily-like AVCN neurons can maintain the temporal-place code of AN fibres and may be modeled by a temporal synchrony detector utilizing the zero-crossings with peak amplitudes (ZCPA) auditory model [1]. PVCN, related to the broadband transients like onsets and other dynamic behaviours, is modeled as the average (mean) discharge rate (MR) derived from an IHC-synapse model consisting of HWR, synaptic adaptation, and a low-pass filter. The DCN, consisting of build-up/pauser neurons, have inhibitory effects similar to lateral inhibition by suppressing firing rates of the neighboring neurons, and thereby improving spectral contrast. This is modelled by two-tone suppression using a temporal companding (compression following by expansion) strategy [7].

Fig. 2 shows the CN computational model proposed by us. It is noted that the IHC-synapse processing is essentially separated from synchrony processing. Three sets of 13 cepstra were generated from the parallel processing of each module. A weighted composite 26-dimensional cepstrum vector was constructed from these by the fusion by early integration of the first 13, 7 and 6 cepstra from synchrony, two-tone suppression and synapse (MR) modules, respectively. An early integration was implemented rather than late integration to reduce processing time, and due to the fact that all the three sub-systems have equal frame rates, and that the reliability weight of each module is taken into account by the number of cepstral coefficients contributed by that sub-system. The emphasis of the MR was kept lowest due to its susceptibility to noise. Two different frequency bands and window lengths in the synchrony and the MR paths were used to enhance the particular segments of the speech. The synchrony filterbank used the frequency range 10-3500 Hz with a maximum derivative window of 80 ms, while the MR filterbank used the frequency range 10-8500 Hz with constant 25 ms hamming windows, each utilizing 16 FIR filters in ERB scale. The frame rates were synchronized at 10 ms.

3. AVCN processing- synchrony detection

The ZCPA [1] is an auditory model used as an ASR front-end and also for speech sound processing in cochlear implants. It utilizes a zero-crossing algorithm for feature extraction from a frequency histogram constructed from the inverse of the time interval between two upward going zero-crossings of the input signal. Fig. 3 shows the schematic of the ZCPA model [1]. If \( s(t) \) is the input speech corrupted with white noise, the ensemble interval histogram output \( y(t,f) \) is obtained by summing the frequency bin weights over all channels and frames,

\[
y(t,f) = \sum_{i=1}^{C} \sum_{k=1}^{D_i} \log(P_{ik} + 1) \delta_{b,k}, \quad 1 < b < R \tag{1}
\]

where \( \delta_b \) is the Kronecker delta, \( C \) the number of cochlear filters, \( D \) the number of zero-crossings in the \( i \)-th channel, \( R \) is
The mean rate captures the very rapidly changing dynamic nature of speech and is more important in characterizing transient sounds, and determines the envelope amplitude corresponding to the average discharge rate response \[9\]. The mean rate output is obtained from the inner hair cell (IHC) synapse model as shown in Fig. 2, consisting of a half-wave rectifier, an adaptation stage and a low-pass filter for synchrony suppression. The output \( Y_M \) for the \( i \)-th filter is obtained from the low-pass smoothing filter (ED) to extract the signal envelope, given by

\[
Y_M(t) = y_{hi}(t) * h_a(t) * h_i(t) * h_{ED}(t), \quad i = 1, \ldots, C, \quad (3)
\]

\( C \) is the number of channels and \( * \) indicates the convolution.

The bending of the cilia attached to the IHC acts as autozeroing half-wave rectifier for the velocity of the motion of the cochlear fluid due to a stimulus. The HWIR in the MR path is implemented as given in \[9\]

\[
y_{hi}(t) = \left\{ \begin{array}{ll}
1 + A \tan^{-1} B x_i, & x_i > 0 \\
e^{-A B x_i}, & x_i \leq 0
\end{array} \right.
\]

where \( x_i \) and \( y_{hi} \) are the input and output in the \( i \)-th channel \((i=1, \ldots, C)\), respectively, and \( A=10 \) and \( B=65 \).

Short-term adaptation improves immunity of the system to stationary noise and also enhances the signal onsets \[2\]. In the mean rate, the synaptic adaptation process is simulated by a high-pass infinite impulse response (IR) filter given in \[11\] as

\[
H_a(z) = \frac{10 \tau f_r (1 - z^{-1})}{(10 \tau + 0.05) + (10 f_r - 0.05) z^{-1}} \quad (4)
\]

where \( \tau=0.25 \) is the short-term adaptation time constant in secs and \( f_r \) is the frame rate equal to 100 Hz.

Neural spikes tend to occur in synchrony with the stimulus, which is reduced at higher frequencies. Loss of synchrony is a low-pass phenomena which was implemented with a first order low-pass filter with a time constant of 0.04 ms as given in \[9\]. The output \( y_{hi} \) was filtered by a first order low-pass smoothing filter (envelope detector, ED) to extract the envelope of the signal. The maximum amplitude in each frame was retained for the \( M \)th and accumulated for all frames, \( m \) to obtain \( Y_M(m,f) \) across the filterbank. Each output \( Y_M(m,f) \) may be considered as the short-term average discharge of all the fibres sensitive to a particular CF, and can be modeled probabilistically. Usually \( Y_M \) is continuous density multimodal with a finite number of Gaussian mixtures. If the number of mixtures is denoted by \( g \), in proportion \( \pi_1, \ldots, \pi_g \) then the p.d.f. of an observation \( Y_M \) can be expressed in finite mixture form as

\[
f(Y_M; \phi) = \sum_{i=1}^{g} \pi_i f(Y_M; \theta_i) \quad (5)
\]

where \( f(Y_M; \theta_i) \) is the p.d.f. corresponding to \( g_i, \theta_i \) denotes the vector of all unknown parameters associated with the parametric forms adopted for the \( g \) component densities, and \( \phi \) denotes a vector of all unknown parameters associated with \( Y_M, \theta_i \). The variance and mean (with \( E \) as expectation) are given by

\[
\text{Var}(Y_M) = E[Y_M^2] - \left[ \int \text{all } Y_M Y_M f(Y_M)dY_M \right]^2 \quad (6)
\]

\[
E(Y_M) = \int \text{all } Y_M Y_M f(Y_M)dY_M \quad (7)
\]
The rate-place representation is more sensitive to background noise than the synchrony response [8]. Fig. 5 shows the mean rate spectrogram for the male utterance ‘one’ in clean condition (upper panel) and with 15 dB SNR white noise (lower panel) with time constants (a) $\tau=0.1$ ms and (b) $\tau=10$ ms of the smoothing low-pass filter. In gray scale, lighter shade indicates higher intensity and darker shade implies lower intensity.

The rate-place representation is more sensitive to background noise than the synchrony response [8]. Fig. 5 shows the mean rate spectrogram for the male utterance ‘one’ in clean condition (upper panel) and with 15 dB SNR white noise (lower panel) with time constants (a) $\tau=0.1$ ms and (b) $\tau=10$ ms of the smoothing low-pass filter. It is observed that in clean condition, the high frequency (higher channels) components are enhanced at the higher time constant of $\tau=10$ ms. Moreover, in 15 dB SNR noise (lower panel), the noise is reduced (within the higher frequency speech segments) at a higher time constant. This is consistent with [2] where it is shown that longer time constants may give better recognition performance.

### 6. HMM Recognition results

Continuous density HMM recognition was performed with isolated digits from the TDigits with 15 states per digit and 5 mixture components per state. A 3-state silence/pause model was inserted at the beginning of each utterance. There were 55 male speakers in the training set (total 1210 utterances) and 30 speaker-independent 31 male speakers in the test set (total 682 utterances). Test speech was corrupted with added Gaussian white noise and non-stationary factory noise. Table 1 summarizes the word recognition rates with our proposed CN processing strategy. It is observed that in stationary white noise, ZCPA_AUD performs better than base ZCPA in clean conditions, but degrades in noise due to the susceptibility of the MR to white noise perturbations. In non-stationary factory noise, the performances of the ZCPA_AUD are improved over the base ZCPA under all signal conditions. The mel-frequency cepstral coefficients (MFCC) with delta features (MFC_DEL) are also compared with ZCPA_AUD, both restricted to 26 dimensions for a fair comparison. The MFCC features were obtained with a 25 ms window length at 10 ms frame rate with 40 triangular filters (13 linear and 27 log). The results indicate that ZCPA_AUD performs equal or better than MFC_DEL in white noise, but degrades in non-stationary factory noise.

### 7. Discussion and conclusions

We propose a feature extraction method based on the processing strategy of the CN. Features were extracted by the parallel processing of three subsections corresponding to the AVCN, DCN and PVCN of the CN, each of which is assigned a particular task of detecting and enhancing the particular speech segments.

### Table 1: HMM recognition rates (%) of ZCPA with ZCPA_AUD and MFC_DEL (26 dimensional) with isolated TDigits.

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>White noise</th>
<th>Factory noise</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZCPA_AUD</td>
<td>ZCPA_DEL</td>
</tr>
<tr>
<td>5</td>
<td>50.0</td>
<td>22.7</td>
</tr>
<tr>
<td>15</td>
<td>77.3</td>
<td>31.8</td>
</tr>
<tr>
<td>30</td>
<td>81.8</td>
<td>86.3</td>
</tr>
<tr>
<td>40</td>
<td>90.9</td>
<td>99.8</td>
</tr>
</tbody>
</table>

The system was evaluated by HMM recognition with features extracted from weighted cepstrum of the three processing using isolated digits. In noisy conditions, improvements, improvements are observed in non-Gaussian real-world noise at all SNRs. The improvements are mainly due to the separation of mean rate from the synchrony path which is consistent with the morphological processing in the CN, and also due to effects of the mean rate synapse processing on the high frequency articulations. However, there is a degradation in white noise at low SNRs due to greater susceptibility of the mean rate to broadband noise. It is further observed that a high time constant of the low-pass filter used for envelope detection may improve system performance in clean and white noise conditions. Some limitations are increased dimensionality of the new extracted features and the computational cost of the overall proposed approach compared with MFCC and the perceptual linear prediction (PLP). Optimizing the processing algorithms using noise-compensated models for the MR processing, and output data fusion by late integration with reliability factors are future research directions.

### 8. References