ACCURATE SPEECH SEGMENTATION BY MIMICKING HUMAN AUDITORY PROCESSING

Sarah King and Mark Hasegawa-Johnson

University of Illinois
Department of Electrical and Computer Engineering
Urbana-Champaign, Illinois

ABSTRACT

This paper addresses the problem of locating phone boundaries without prior knowledge of the text of an utterance. A biomimetic model of human auditory processing is used to calculate the neural features of frequency synchrony and average signal level. Frequency synchrony and average signal level are used as input to a two-layered support vector machine (SVM)-based system to detect phone boundaries. Phone boundaries are detected with 87.0% precision and 84.8% recall when the automatic segmentation system has no prior knowledge of the phone sequence in the utterance.

***Index Terms*—** Automatic segmentation, auditory modeling, frequency synchrony, average signal level

1. INTRODUCTION

A "chicken and egg" dichotomy currently exists between automatic speech segmentation and automatic speech recognition. On one hand, speech recognition systems require accurately segmented training transcriptions in order to recognize the phone sequence. On the other hand, speech segmentation systems require accurate knowledge of the phone sequence in order to split the speech signal into its phonetic components. The requirement that an accurate transcription must be available in order to segment speech is especially pernicious in the study of under-resourced languages and dialects, for which accurate pronunciation dictionaries may not exist. Improved training of ASR based on small training corpora requires improved segmentation of the available training data.

The human brain is the fastest, most accurate computer with respect to cognitive tasks such as speech recognition and speech segmentation. In the human auditory brainstem, the speech signal is segmented before phones and words are recognized. The octopus cells in the cochlear nucleus detect speech onsets [1] by detecting coincident firing of auditory nerve fibers [2]. Subsequent processing in the lateral lemniscus may also aid in the detection of sound onsets [3]. The recognition of phones and words does not take place until the signal reaches the auditory cortex and beyond [4].

This work is supported by NSF grant IIS 0807329.

The human brain detects phone boundaries by first dividing the speech signal into several thousand frequency channels [5]. Octopus cells detect synchrony between subsets of these frequency channels [2]. Neurons in the lateral lemniscus may detect the rate at which synchrony occurs[1]. Multipolar neurons in the cochlear nucleus also respond to signal onsets [6] and encode signal level [7]. Signal level and rate of synchrony are two important neural cues that humans use for speech signal segmentation. Are these cues useful to computers?

This paper presents an automatic speech segmentation system that estimates phone boundaries using the average signal level and rate of synchrony. Average signal level and rate of synchrony are calculated by a biomimetic model of the human auditory system. The system presented in this paper can accurately mark phone boundaries without prior knowledge of the phone sequence that contained in the utterance.

2. PREVIOUS WORK

2.1. Auditory features

The Mel-frequency cepstral coefficient (MFCC) [9] and the perceptual linear prediction (PLP) coefficient [10] are calculated by warping the frequency axis of the spectrum along the Mel and Bark frequency scales, respectively. This warping mimics the frequency resolution along the basilar membrane (BM). The MFCC captures cochlear compression effects — suppressing spectral variation in the higher frequency bands — by taking the logarithm of the spectrum. The PLP attempts to capture human frequency sensitivity and to simulate the relationship between sound intensity and perceived loudness. Both MFCCs and PLPs are commonly used features for speech segmentation and speech recognition.

The Lyon [11, 12] and Seneff [13] auditory models emulate cochlear filtering using a bandpass filter bank. Both models emulate the transmission of cochlear filter outputs to nerve

[1]The lateral lemniscus is known to calculate the rate of sound position change [8]. It is likely, though to the authors' knowledge it has not been shown, that the nucleus is also responsible for tracking the first derivatives of other functions, such as synchrony.

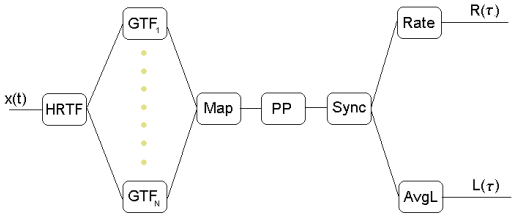

Fig. 1. The auditory model used for acoustic feature creation. The signal x(t) is filtered by a head-related transfer function (HRTF) and then separated into components using bandpass gammatone filters (GTF). GTF output is used to create a tonotopic map (MAP). A peak-picker (PP) is used to locate local maxima. Synchrony (Sync) is calculated between the peaks. The synchronous information is used to calculate the rate of synchrony (Rate) and average signal level (AvgL).

responses. Additionally, the Seneff auditory model emulates the synchrony detection that occurs in the cochlear nucleus. Both models can be used for speech analysis.

2.2. Automatic speech segmentation

When neither phone sequence nor phone boundaries are known, simultaneous phone recognition and automatic segmentation on TIMIT yields error rates between between 45% and 27% [14], e.g., Duskan and Rabiner segment TIMIT utterances with 84.6% recall and with a precision of 75.0% [15]. Prior knowledge of the phone sequence reduces the segmentation error [16, 17].

3. AUDITORY FEATURE CREATION

Figure 1 shows the auditory model used to compute average signal level and rate of synchrony. Components of this model have been previously reported, e.g., the model can be used to detect acoustic landmarks in a manner robust to changes in the bandwidth and noise level of the speech signal [18]. This section describes key features of the model.

Before a sound $x(t)$ reaches the cochlea, it is filtered by the head and outer ear. This head-outer ear filter is referred to as the head-related transfer function (HRTF). The auditory model uses the finite impulse response (FIR) HRTF measured by Tidemann [19].

We model the frequency analysis performed by the basilar membrane (BM) using a bank of $N = 2760$ parallel gammatone filters designed according to [20]. The center frequencies of the filters range from $f_1 = 60$ Hz – f_{2760} Hz and are spaced according to the equivalent rectangular bandwidth (ERB) [21] scale. There are 100 filters per critical band. Filterbank outputs are concatenated along the frequency axis to form a tonotopic map. An example of such a map for the

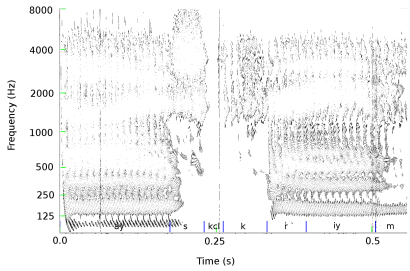

Fig. 2. A topographic map of the bandpass filter output for the utterance " ice cream" spoken by subject MADC0 from the TIMIT corpus. Darker regions are higher in amplitude.

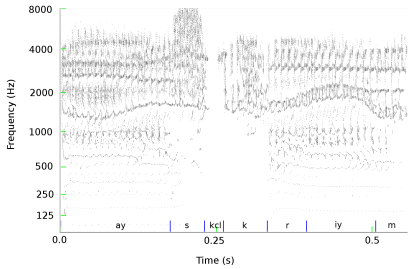

Fig. 3. Local maxima from the tonotopic map of the sentence "ice cream" produced by the speaker MADC0 from the TIMIT corpus shown in Figure 2. Darker colors indicate higher amplitude.

words "ice cream" is shown in Figure 2. We collect the frequency, timing, and intensity level information at local maxima in the tonotopic map to both mimic the inner hair cells and auditory nerve, and to create a sparse representation of the signal. An example of this representation is shown in Figure 3.

The intensity level is calculated from the output of the gammatone filters.

$$I(t, f_m) = 20 \log_{10} \frac{y_m(t)}{Y_{ref}} \quad (1)$$

Here, $I(t, f_m)$ is the intensity level in decibels in the m^{th} frequency band at time t, $y_m(t)$ is the observed output at time t from the m^{th} filter given that a maximum has been found, and Y_{ref} is the minimum detection threshold.

Level, frequency, and timing information are stored as a sparse binary third order tensor A. An individual entry in A is referenced by its time, frequency, and intensity level values and $A(t, f, i) \in \{0, 1\}$. A value $A(t, f, i) = 1$ indicates that we have detected a signal component at f Hz with level i dB at time t.

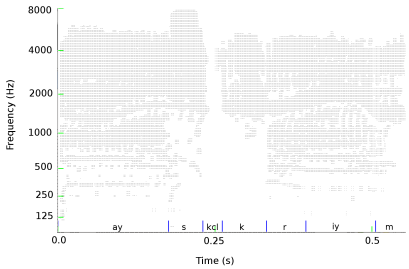

Fig. 4. Synchrony detection on the data in Figure 3 for the words "ice cream" produced by the speaker MADC0 from the TIMIT corpus. A dot indicates that synchrony between frequency bands was detected at time t.

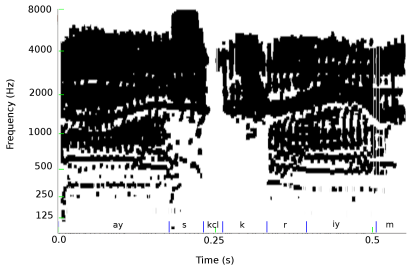

Fig. 5. Rate of synchrony for the data in Figure 4 for the words "ice cream" produced by the speaker MADC0 from the TIMIT corpus. Darker colored regions have a higher rate.

Synchrony is detected using the logical union ($\cup_i$) of the binary variables $A(t, f, i)$ over different values of i, and summing over a time-frequency window of duration T_w and over F_w frequency bands. In other words,

$$S_w(t,f) = \sum_{\tau=0}^{T_w-1} \sum_{\phi=0}^{F_w-1} \cup_{i=I_{min}}^{I_{max}} A(t-\tau, f-\phi, i) \qquad (2)$$

and

$$O_w(t,f) = \begin{cases} 1 & S_w(t,f) > \rho \\ 0 & \text{otherwise} \end{cases}$$

where ρ is the minimum number of data points in a window w required for synchrony to be detected. The optimum window size was determined experimentally to be 3 ms by 0.6 ERB with an optimum firing threshold of $\rho = 2$. The frequency step is 0.2 ERB. The time step is 1 ms. This calculation mimics the octopus cells in the cochlear nucleus. Synchrony detector output is shown in Figure 4 for the words "ice cream."

The rate of synchrony is determined as follows

$$R_{O_w(t,f)} = \frac{1}{\tau(O_w(t_m,f)) - \tau(O_w(t_n,f))} \qquad (3)$$

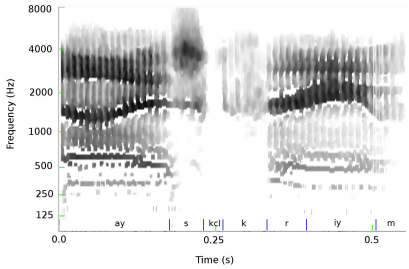

Fig. 6. Average signal level for the data in Figure 4 for the words "ice cream" produced by the speaker MADC0 from the TIMIT corpus. Darker colored regions have a higher level.

where $\tau(O_w(t,f)) = t$, and $O_w(t_m,f)$ and $O_w(t_n,f)$ are two chronologically ordered, nonzero instances of synchronous activation, i.e., $t_m > t_n$. This calculation mimics processing in the lateral lemniscus. The rate of synchrony for the words " ice cream" is shown in Figure 5.

The average spectral level is calculated as follows

$$L_w(t,f) = \frac{\sum_{\tau=0}^{T_w-1} \sum_{\phi=0}^{F_w-1} I(t-\tau, f-\phi)}{T_w F_w} \qquad (4)$$

Level is summed from all active points in a time-frequency window of duration T_w and over F_w frequency bands. The window size used to calculate average signal level is 3 ms by 0.6 ERB. This mimics the function of multipolar cells in the cochlear nucleus. The average signal level of the words "ice cream" is shown in Figure 6.

4. AUTOMATIC SPEECH SEGMENTATION

Six radial basis function (RBF) support vector machines (SVMs) were trained using the TIMIT corpus to detect the phone boundaries where one edge of the boundary belongs to one of six different broad phonetic classes (stop closures, fricatives, glides and liquids, nasals, stop releases, and vowels). The six RBF SVMs feed into a single SVM trained to detect transition regions between phones.

The phone boundary detection SVMs were trained using both MFCCs and the neural rate and level (NRL) features described in this paper. Each frame of MFCCs contained 39 coefficients (including deltas, acceleration, and energy). MFCCs were calculated using a 25 ms window with a time-step of 5 ms. NRL features were calculated every millisecond to match the maximal firing rate of the octopus cells in the cochlear nucleus. The NRL feature vector is a 276 dimensional vector composed of 138 instances of $O_w(t,f)$ and 138 instances of $L_w(t,f)$ for the window w at time t.

The training and testing input for the phone boundary detection SVMs consists of feature vectors $\vec{x}_t$ containing 11

Table 1. Support vector machine frame classification accuracy for stop closures (SC), fricatives (F), glides and liquids (GL), nasals (N), stop releases (SR), and vowels (V). Also shown is the frame classification accuracy for the transition/steady-state region (R) classification SVM. Results are shown for SVMs using both MFCCs and NRLs.

	SC	F	GL	N	SR	V	R
MFCC	95.3	93.2	91.9	96.3	95.5	96.7	83.7
NRL	94.6	94.1	91.9	93.2	95.4	97.2	83.0

concatenated acoustic feature frames. The first frame in $\vec{x}_t$ was sampled at 50 ms before the phone boundary, the 6th frame was sampled at the boundary time t, and the 11th frame was sampled at 50 ms after the phone boundary; i.e., $\vec{x}_t \equiv [\vec{y}_{t-50}, \ldots, \vec{y}_t, \ldots, \vec{y}_{t+50}]$. In other words, $\vec{x}_t$ is created by concatenating n acoustic feature frames on both sides of the frame corresponding to the phone boundary $\vec{y}_t$, where the time step between frames is 10 ms and the total number of concatenated frames in $\vec{x}_t$ is $2n + 1$. The vector $\vec{y}_t$ contained either MFCCs or NRLs

The phone boundary detection SVMs were used to generate a discriminant function for every frame in every file in the training corpus. The discriminant function was then smoothed. The smoothed values at each time t were concatenated into feature vectors $\vec{d}_t$. These discriminant feature vectors were used to train an SVM that classified frames as either transitional regions between phones or as steady-state regions. The input to the SVM classifier consisted of feature vectors $\vec{D}_t = [\vec{d}_{t-30} \ldots \vec{d}_t \ldots \vec{d}_{t+30}]$.

The training corpus consisted of the SX TIMIT audio files. The test set consisted of the SI TIMIT audio files. A total of 10000 training tokens (5000 boundary tokens and 5000 non-boundary tokens) were extracted from the training data. A total of 8000 tokens (4000 boundary tokens and 4000 non-boundary tokens) were extracted from the test set. Transition classification SVMs were trained on 7500 transition tokens and 7500 steady-state tokens. No tokens overlap between either of the training and testing sets, respectively.

5. RESULTS

Phonetic boundary detection and transition classification SVM accuracies are shown in Table 1. Precision, recall, and F-score of the MFCC and NRL-based segmentation systems are given in Table 2. A boundary is labeled correctly if the generated label is within 20 ms of the manually labeled time.

The frame classification SVM was used to generate a discriminant function for each utterance in TIMIT. An example of such a discriminant function is shown in Figure 7. In the figure, the discriminant function is positive whenever a frame is classified as transitional. The phone boundary is determined by finding the start and end times of each transitional region and taking the average.

Table 2. Precision, recall, and F-scores of the automatic phone boundary detectors for the MFCC and NRL-based systems. Also shown are the precision, recall, and F-score from [15] for the same task (DR). Precision, recall, and F-score are calculated based on the numbers of detected, missed, and inserted tokens reported in [15].

	Precision	Recall	F-score
MFCC	20.8	16.5	18.4
NRL	87.0	84.8	85.9
DR	75.0	84.6	79.5

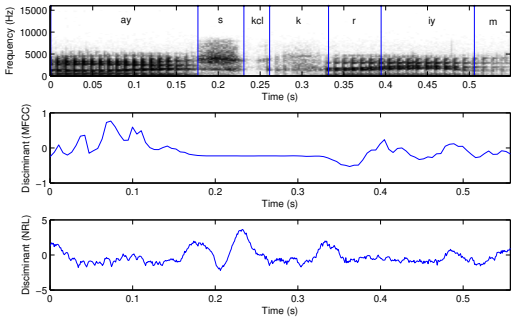

Fig. 7. **Top.** A spectrogram of the words "ice cream." **Middle.** The discriminant generated for the words using the MFCC-based system. **Bottom.** The discriminant generated for the words using the NRL-based system.

Despite the success of the MFCC-based SVMs (See Table 1), the MFCC-based SVM discriminant function does not convey an accurate representation of the phone boundaries (See Figure 7). Key characteristics of the MFCC vector may be similar between some boundary and non-boundary tokens. This may cause the phonetic frame classifiers to generate a large number of both false positives and false negatives. These errors would then propagate to the second layer of the system.

6. CONCLUSION

This paper presents a system that can accurately locate phone boundaries without knowing the phone sequence. Such a system is useful for the study of under-resourced languages, and of other ASR tasks using small training corpora, for which accurate pre-segmentation of the training corpus may improve the ability of the ASR to learn good phone models. Future research will focus on the use of these segmentation boundaries to improve training and test accuracy of large vocabulary speech recognition.

7. REFERENCES

[1] W. Hemmert, M. Holmberg, and U. Ramacher, "Temporal sound processing by cochlear nucleus octopus neurons," in *ICANN*, 2005.

[2] D. Oertel, R. Bal, S. Gardner, P. Smigh, and P. Joris, "Detection of synchrony in the activity of auditory nerve fibers by octopus cells in the mammalian cochlear nucleus," *PNAS*, vol. 97, no. 22, 2000.

[3] E. Covey and J. Casseday, "Timing in the auditory system of the bat," *Annu. Rev. Physiol.*, vol. 61, 1999.

[4] R. Zatorre, P. Belin, and V. Penhune, "Structure and function of auditory cortex: Music and speech," *Trends in Cognative Science*, vol. 6, no. 1, 2002.

[5] C. D. Geisler, *From Sound To Synapse: Physiology Of The Mammilian Ear*, Oxford University Press, Oxford, NY, 1998.

[6] I. Winter, A. Palmer, L. Wiegrebe, and R. Patterson, "Temporalcoding of the pitch of complexsounds by presumed multipolar cells in the ventral cochlear nucleus," *Speech Communication*, vol. 41, no. 1, 2003.

[7] W. Rhode and P. Smith, "Encoding timing and intensity in the ventral cochlear nucleus of the cat," *Journal of Neurophysiology*, vol. 56, 1986.

[8] R. Burger and G. Pollak, "Reversible inactivation of the dorsal nucleus of the lateral lemniscus reveals its role in the processing of multiple sound sources in the inferior colliculus of bats," *The Journal of Neuroscience*, vol. 21, no. 13, 2001.

[9] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continously spoken sentences," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, 1980.

[10] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *JASA*, vol. 87, no. 4, 1990.

[11] R. F. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *ICASSP*, 1982.

[12] R. F. Lyon, "Experiments with a computational model of the cochlea," in *ICASSP*, 1986, pp. 1975–1978.

[13] S. Seneff, "A computational model for the peripheral auditory system: Application to speech recognition research," in *ICASSP*, 1986, pp. 1986–1986.

[14] F. Sha and L. Saul, "Large margin hidden markov models for automatic speech recognition," *Advances in neural information processing systems*, vol. 19, 2007.

[15] S. Duscan and L. Rabiner, "On relation between maximum spectral transition position and phone boundaries," in *ICSLP*, 1993.

[16] A. Ljolje and M. Riley, "Automatic segmentation and labeling of speech," in *ICASSP*, 1991.

[17] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on hidden Markov models," *Speech Communication*, vol. 12, 1993.

[18] S. King and M. Hasegawa-Johnson, "Detection of acoustic-phonetic landmarks in mismatched conditions using a biomimetic model of human auditory processing," in *COLING*, 2012.

[19] J. Tidemann, "Characterization of the head-related transfer function using chirp and maximum length excitation signals," M.S. thesis, University of Illinois, 2011.

[20] R. Patterson and J. Holdsworth, "A functional model of neural activity patterns and auditory images," *Advances in Speech, Hearing, and Language Processing*, vol. 3, 1996.

[21] B. Moore and B. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *JASA*, vol. 74, no. 3, September 1983.