Automatic Phoneme Segmentation Using Auditory Attention Features

Ozlem Kalinli

Sony Computer Entertainment US R&D, Foster City, California, USA
ozlem_kalinli@playstation.sony.com

Abstract

Segmentation of speech into phonemes is beneficial for many spoken language processing applications. Here, a novel method which uses auditory attention features for detecting phoneme boundaries from acoustic signal is proposed. The auditory attention model can successfully detect salient audio events/sounds in an acoustic scene by capturing changes that make such salient events perceptually different than their neighbours. Therefore, it naturally offers an effective solution for segmentation task. The proposed phoneme segmentation method does not require transcription or acoustic models of phonemes. When evaluated on TIMIT, the proposed method is shown to successfully predict phoneme boundaries and outperform the recently published text-independent phoneme segmentation methods [1, 2].

Index Terms: speech segmentation, phoneme boundary detection, auditory attention model.

1. Introduction

Segmentation of continuous speech into phonemes is beneficial for many applications including speech analysis, automatic speech recognition (ASR) and speech synthesis. However, manually determining phonetic transcriptions and segmentations requires expert knowledge and this process is laborious and expensive for large databases. Thus, many automatic segmentation and labeling methods have been proposed in the past to tackle this problem [1, 2, 3, 4, 5].

Phoneme segmentation methods can be grouped in two main categories. The first group of methods requires transcriptions and acoustic models of phonemes, and segmentation task is simplified to the HMM-based forced-alignment of speech with its transcription [3]. One of the drawbacks of this approach is that it assumes the availability of the phonetic transcription. When the transcription is not available, one may consider using a phoneme recognizer for the segmentation. However, speech recognition techniques like HMMs cannot place phone boundaries accurately since they are optimized for the correct identification of the phone sequence [4]. Also, when there is mismatch between the trained acoustic models and the test utterance due to noise conditions, speaking style, and speaker traits (e.g., adult vs. kids), the segmentation performance can drastically degrade.

The second group of methods does not require any prior knowledge of transcription or acoustic models of phonemes. The method proposed here falls under this category. Most of the approaches in this category focused on change point detection for phoneme segmentation. For example, [1] assumed that the maximum spectral transition positions correspond to phoneme boundaries and [5] used maximum margin clustering to locate phoneme boundaries. Differently, [2] proposed a probabilistic approach using an objective function derived from information rate distortion theory. MFCC features were used for phoneme segmentation in [1, 2, 5].

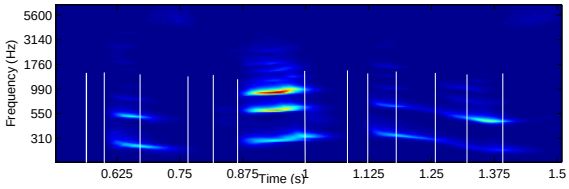

Figure 1: A spectrum of a sample speech segment with transcription "his captain was" containing phonemes: /hh/, /ih/,/z/, /kcl/, /k/, /ae/, /pcl/, /t/, /ix/, /n/, /w/, /ax/, and /s/. Vertical white bars indicate approximate phoneme boundaries.

In our previous work, a bottom-up saliency-driven auditory attention model was proposed and the model could successfully detect salient audio events/sounds in an acoustic scene by capturing changes that make such salient events perceptually different than their neighbours [6, 7]. Recently, auditory attention features were successfully used to find boundaries between syllable nuclei and consonants surrounding it [8]. Hence, the auditory attention model is found to be very effective for change point detection in different tasks.

In this study, a novel method that uses auditory attention cues for phoneme segmentation of speech is proposed. Our motivation for the proposed method is as follows: in a speech spectrum, one can usually see edges and local discontinuities around phoneme boundaries; especially around vowels since they exhibit high energy and clear formant structure. For example, in Fig 1, the spectrum of a speech segment which is transcribed as "his captain was" is shown together with approximate phoneme boundaries. In the spectrum, one can visually observe some of these boundaries that correspond to phoneme boundaries such as the boundaries for vowels ih, ae, ix etc. Hence, we believe that by detecting the relevant oriented edges and discontinuities in the auditory spectrum; i.e. as done visually, phoneme segments and/or boundaries in speech can be located.

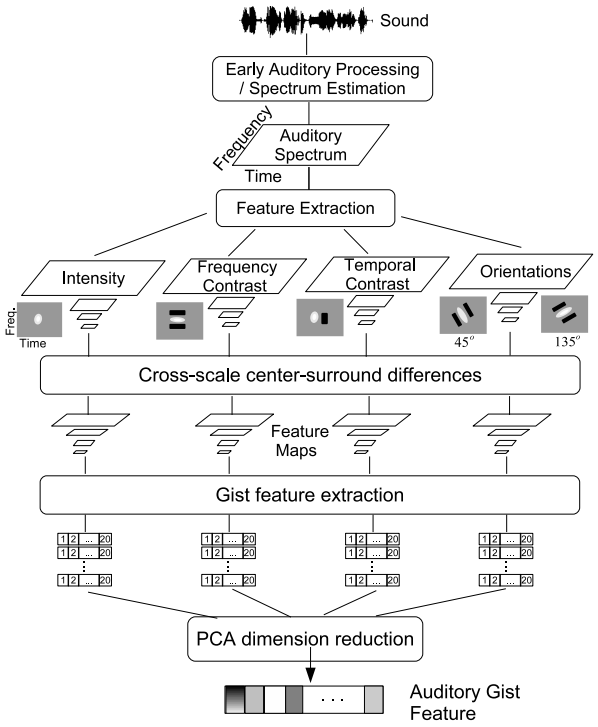

Figure 2: Auditory Attention Model and Gist Extraction

In the auditory attention model, the auditory spectrum is analogous to an image of a scene in vision and contrast features are extracted from the spectrum in multi-scales using 2D spectro-temporal receptive filters. The extracted features are tuned to different local oriented edges: i.e., frequency contrast features are tuned to local horizontally oriented edges, which are good for detecting and capturing formants and their changes [8]. Next, low level auditory gist features are obtained and a neural network is used to discover the relevant oriented edges and to learn the mapping between the gist features and phoneme boundaries.

The rest of the paper is organized as follows. The auditory attention model and features are explained in Section 2, which is followed by experiments and a discussion in Section 3-4. A conclusion is presented in Section 5.

2. Auditory Attention Model

The block diagram of the auditory attention model is shown in Fig 2. The model is biologically inspired and hence mimics the processing stages in the human auditory system. First, the auditory spectrum of the input sound is computed based on early stages of the human auditory system, which consists of cochlear filtering, inner hair cell, and lateral inhibitory stages, mimicking the process from basilar membrane to the cochlear nucleus in the auditory system [6]. The cochlear filtering is implemented using a bank of 128 overlapping constant-Q asymmetric band-pass filters with center frequencies that are uniformly distributed along a logarithmic frequency axis. For analysis, audio frames of 25 milliseconds (ms) with 6.25 ms shift are used, i.e. each 6.25 ms audio frame is represented by a 128 dimensional vector to have enough resolution for phoneme boundary detection.

In the next stage, multi-scale features, which consist of *intensity* (I), *frequency contrast* (F), *temporal contrast* (T), and *orientation* (O_θ) with $\theta = \{45^o, 135^o\}$, are extracted from the auditory spectrum based on the processing stages in the central auditory system [6, 9]. These features are extracted using 2D spectro-temporal receptive filters mimicking the analysis stages in the primary auditory cortex. Each of the receptive filters (RF) simulated for feature extraction is illustrated with gray scaled images in Fig 2 next to its corresponding feature. The excitation phase and inhibition phase are shown with white and black color, respectively. For example, F filter corresponds to receptive fields in the primary auditory cortex with an excitatory phase and simultaneous symmetric inhibitory side bands. Each of these filters is capable of detecting and capturing certain changes in signal characteristics. For example, F is capable of detecting and capturing changes along the spectral axis, whereas O_θ is capable of capturing and detecting moving ripples (i.e. raising and falling curves). One important point is that in the model contrast features are computed, which is crucial for change point detection and segmentation.

The RF for I has only an excitation phase and is implemented using a 2D Gaussian kernel. The RF for F, T, O_θ is implemented using 2D Gabor filters with angles 0^o, 90^o, $\{45^o, 135^o\}$, respectively. The multi-scale features are obtained using a dyadic pyramid: the input spectrum is filtered and decimated by a factor of two, and this is repeated. Finally, eight scales are created (if the window is larger than 0.8 ms; otherwise there are fewer scales), yielding size reduction factors ranging from 1:1 (scale 1) to 1:128 (scale 8). For details of the feature extraction and filters, one may refer to [6, 9].

After multi-scale features are obtained, the "center-surround" differences are computed by comparing a "center" fine scale c with "surround" coarser scale s, yielding a feature map $\mathcal{M}(c, s)$:

$$\mathcal{M}(c, s) = |\mathcal{M}(c) \ominus \mathcal{M}(s)|, \;\; \mathcal{M}\epsilon\{I, F, T, O_\theta\} \quad (1)$$

The center-surround operation mimics the properties of local cortical inhibition and detects local temporal and spatial discontinuities. The across scale subtraction ($\ominus$) between two scales is computed by interpolation to the finer scale and point-wise subtraction. Here, $c = \{2, 3, 4\}$, $s = c + \delta$ with $\delta\epsilon\{3, 4\}$, which results in 30 feature maps when there are 8 scales.

Next, an "auditory gist" vector is extracted from the feature maps of I, F, T,O_θ such that it covers the whole scene at low resolution. To do that, each feature map is divided into m-by-n grid of sub-regions and mean of each sub-region is computed to capture the overall properties of the map. For a feature map $\mathcal{M}_i$ with height h and width w, the computation of feature can be written as:

$$G_i^{k,l} = \frac{mn}{wh} \sum_{u=\frac{kw}{n}}^{\frac{(k+1)w}{n}-1} \sum_{v=\frac{lh}{m}}^{\frac{(l+1)h}{m}-1} \mathcal{M}_i(u,v), \qquad (2)$$

where $k = \{0, \cdots, n-1\}, l = \{0, \cdots, m-1\}$, and feature map index $i = \{1, \cdots, 30\}$. An example of gist feature extraction with $m = 4, n = 5$ is shown in Fig 2, where a $4 \times 5 = 20$ dimensional vector represents a feature map. After extracting a gist vector from each feature map, we obtain the cumulative gist vector by augmenting them. Then, principal component analysis (PCA) is used to remove redundancy and to reduce the dimension.

3. Experiments and Results

TIMIT database is used in automatic phoneme boundary detection experiments since it contains phone boundaries manually determined by experts. We used the official train and test splits, which contain 1344 and 3696 utterances with total of 50337 and 139214 between-phone boundaries, respectively. In this section, phoneme boundary detection results at frame-level for varying window duration W, an analysis of auditory attention features, and a comparison of phoneme segmentation results with earlier work will be presented.

In the experiments, a 3-layer artificial neural network (ANN) is used to learn the mapping between the auditory gist features and phoneme boundaries. ANN has D inputs, $(D + N)/2$ hidden nodes and N output nodes, where D is the length of auditory gist vector after PCA dimension reduction when 95% of the variance is retained, and $N = 2$; i.e. boundary vs. non-boundary.

The auditory gist features are estimated every 12.5 ms using a window of duration W that centers on the current frame to capture the context. A 12.5 ms error margin is allowed for detected phoneme boundaries. For example, if there is a reference boundary at 130 ms, the auditory gist features corresponding to the frames at 125 ms and 137.5 ms are both labeled as a boundary in the training since there is no exact frame which corresponds to manual label. During evaluation, frame/frames detected as a boundary within 12.5 ms window of a reference boundary is/are accepted correct. For the above example, detecting a boundary for either frames located at 125 ms or 137.5 ms is accepted correct, when there is a reference phoneme boundary at 130 ms. The excessive detected boundaries are counted as insertions and having no detected boundary for a reference one is counted as deletion.

First, the role of window duration W is investigated in the experiments by varying duration from 62.5 ms, which is approximately mean phoneme duration ($\mu = 76$ ms), up to 200 ms ($\approx 3 \times \mu$) to analyze the effect of neighbouring left and right context on the performance. The grid size determines the temporal and spectral resolution. Different grid sizes are evaluated for auditory gist extraction for varying temporal and spectral resolution. It was found that a grid size of 16-by-10 performs well in this task with a reasonable feature dimension. In Table 1, the

Table 1: Phoneme Boundary Detection Results at Frame-Level for Varying Window Duration

W(ms)	D	Ac	Pr	Re	Fs
62.5	47	86.49	77.84	66.99	72.0
125	**77**	**86.75**	**77.67**	**67.33**	**72.13**
200	117	85.37	75.37	63.05	68.67

Table 2: Phoneme Boundary Detection Results at Frame-Level for Individual Feature with $W = 125$ ms

Feat.	D	Ac	Pr	Re	Fs
I	38	83.90	74.56	56.34	64.18
F	29	81.43	74.97	42.78	54.48
T	47	86.72	77.12	68.68	72.65
O	32	80.0	72.44	38.29	50.10
IFTO	**77**	**86.75**	**77.67**	**67.33**	**72.13**

frame-level phoneme boundary detection results in terms of Accuracy (Ac), Precision (Pe), Recall (Re), and F-score (Fs), for varying window duration are presented together with the corresponding auditory gist dimension D. The boundary detection performance is lower for longer window duration since it causes missing phoneme boundaries; i.e. when $W = 200$ ms recall is lower. The best performance achieved is 86.75% phoneme boundary detection accuracy at frame-level with $W = 125$ ms.

Second, the contribution of each feature I, F, T, O in the attention model is presented in Table 2 for $W = 125$ ms. All of the features individually perform well above the chance level, which is 69.8% (obtained by labeling all frames with the majority class). The most informative feature about phoneme boundary detection is *temporal contrast (T)*, which achieves 86.72% phoneme boundary detection accuracy at frame-level. However, the highest accuracy is achieved with the combined features $IFTO$. The high performance achieved with temporal contrast features can be attributed to the fact that they are detecting temporal changes in the auditory spectrum.

We cannot directly compare our results with the results reported in the literature due to differences in the parameters, evaluation metrics, data sets used in the experiments, etc. In the literature, the work on phoneme segmentation has focused on detection at segment/phoneme level whereas here a more detailed frame-level results are presented. For comparison with the recently published work in the literature, experiments are also conducted with 10 ms frame shift and phoneme level results are obtained as in [1, 2]. For each frame, the ANN returns a value between $[0, 1]$, which can be thought as $P(B|f)$, the posterior probability of a frame being a phoneme boundary, B, given auditory gist features, f. The ANN output score is used for generating a one-dimensional curve as a function of time, $P_t(B|f)$ and a peak search is performed on the curve to locate local maxima. Finally, peaks that are larger than a threshold are used to locate phoneme boundaries.

Figure 3 presents results for the sample speech segment shown in Fig 1. The first plot at the top dis-

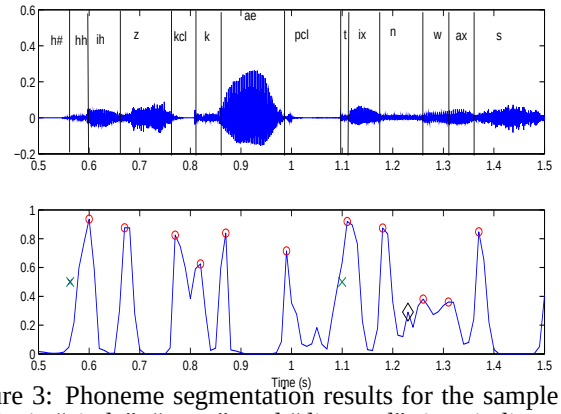

Figure 3: Phoneme segmentation results for the sample in Fig 1. “circle”, “cross”, and “diamond” signs indicate matched, missed, and inserted boundaries, respectively.

Table 3: Comparison of Phoneme Segmentation Methods

Method	Re	Pr	Fs
Dusan et al [1]	75.2	72.73	73.94
Quiao et al [2]	77.5	78.76	78.13
Attention Cues	**81.77**	**84.32**	**83.02**

plays speech waveform with manually placed phoneme boundaries. The second plot displays $P_t(B|f)$. When the threshold was set to 0.2, the method detects most of phoneme boundaries within 20 ms of reference boundaries. However, it misses the boundaries between silence and /hh/, stop closure /pcl/ and /t/, and inserts a false boundary at 1.23 s.

For scoring, a time-alignment between the detected phoneme boundaries and the reference ones is used. However, as done in [1], first, manual phoneme boundaries are converted to the closest adjacent frame positions since there is not always an exact corresponding frame to a manual boundary due to the frame shift size. Then, if a peak is detected within 20 ms window of a reference phoneme boundary, it is accepted as correct. Here, no peak could validate more than one reference phoneme boundary; i.e. in Fig. 3, the boundary between /pcl/ and /t/ was considered as missed even though the peak at 1.11 s is within 20 ms. The excessive detected peaks are counted as insertions, and having no detected peak for a reference phoneme boundary is counted as a deletion.

Phoneme segmentation results are given in Table 3 along with the state-of-the art (to the best of our knowledge) results reported in [1, 2] on TIMIT. The auditory attention features can detect 81.77% of phoneme boundaries with 84.32% precision. The results from Table 3 show that the proposed method with auditory attention features performs better than [1, 2].

4. Discussion

In this section, we compare three variables, namely window duration, grid size, and features, analysed in this study for phoneme segmentation and in [8] for syllable segmentation. The best performance achieved in syllable segmentation and phoneme segmentation was with 400 ms and 125 ms windows, respectively. It is interesting to note that in both tasks,the best performing window is approximately twice the mean duration of segment, where a segment is syllable or phoneme. This indicates that left and right neighbouring context helps in segmentation task. Second, phoneme segmentation requires a larger grid size (16-by-10) compared to syllable segmentation (4-by-10) which indicates that phoneme segmentation needs higher resolution as one expects. Finally, an analysis of attention features showed that the most informative feature is F for syllable segmentation and T for phoneme segmentation. One possible explanation for this is that frequency contrast feature is more discriminative for finding boundaries between vowels and consonants surrounding them, whereas temporal changes in the spectrum are more informative for finding boundaries between phonemes.

5. Conclusion and Future Work

In this paper, biologically inspired auditory attention cues are proposed for phoneme segmentation of continuous speech. A neural network is used to learn the mapping between phoneme boundaries and auditory attention features. The proposed method achieves 86.8% phoneme boundary detection accuracy at frame-level when tested on TIMIT. At phoneme level, it is shown that the proposed method outperforms the recently published text-independent phoneme segmentation methods in [1, 2].

During error analysis, it is found that most of the missed phone boundaries occur during transitions from stops to vowels. As part of future work, we will investigate using phone classes as side information during segmentation to obtain further improvement. We also plan to conduct experiments in other languages, speaking styles, and noise conditions.

6. Acknowledgement

The author would like to thank Dr. Ruxin Chen at SCEA for valuable discussions.

7. References

[1] S. Dusan and L. Rabiner, “On the relation between maximum spectral transition positions and phone boundaries,” in *Proc. of ICSLP*, 2006.

[2] Y. Qiao, N. Shimomura, and N. Minematsu, “Unsupervised optimal phoneme segmentation: objectives, algorithm and comparisons,” in *Proc. of ICASSP*, 2008.

[3] F. Brugnara, D. Falavigna, and M. Omologo, “Automatic segmentation and labeling of speech based on hidden markov models,” *Speech Communication*, vol. 12, no. 4, pp. 357–370, 1993.

[4] A. Sethy and S. S. Narayanan, “Refined speech segmentation for concatenative speech synthesis,” in *Proc. of ICSLP*, 2002.

[5] Y. Estevan, V. Wan, and O. Scharenborg, “Finding maximum margin segments in speech,” in *Proc. of ICASSP*, 2007.

[6] O. Kalinli and S. Narayanan, “A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech,” in *Proc. of Interspeech*, 2007.

[7] O. Kalinli, S. Sundaram, and S. Narayanan, “Saliency-driven unstructured acoustic scene classification using latent perceptual indexing,” in *Proc. of MMSP*, 2009.

[8] O. Kalinli, “Syllable segmentation of continuous speech using auditory attention cues,” in *Proc. of Interspeech*, 2011.

[9] O. Kalinli and S. Narayanan, “Prominence Detection Using Auditory Attention Cues and Task-Dependent High Level Information,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, pp. 1009–1024, 2009.