Combination of Auditory Attention Features with Phone Posteriors for Better Automatic Phoneme Segmentation

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. Previously, a novel method which employs auditory attention features for detecting phoneme boundaries from acoustic signal was proposed in [1] outperforming [2, 3]. In this paper, we propose to use phone posterior features, which are obtained from a Deep Belief Network (DBN) based phoneme recognition system, along with attention features since they provide complementary information. When evaluated on TIMIT corpus, the proposed method is shown to successfully predict phoneme boundaries and outperform the recently published text-independent phoneme segmentation methods. Also, the combination of attention features with posterior features yield more than 30% relative improvement in F-measure over the system which used only attention features.

Index Terms: phoneme segmentation, boundary detection, auditory attention model, deep belief networks.

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 [2, 3, 4, 5, 6].

Phoneme segmentation methods can be grouped into two main categories. The first group of methods requires transcriptions and acoustic models of phonemes, and segmentation task is simplified to an HMM-based forced-alignment of speech with its transcription [4]. 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 [5] rather than the correct detection of boundaries.

The second group of methods does not require any prior knowledge such as transcription or acoustic models of phonemes [2, 3, 6]. Most of the approaches in this category focused on change point detection for phoneme segmentation. For example, [2] assumed that the maximum spectral transition positions correspond to phoneme boundaries and [6] used maximum margin clustering to locate phoneme boundaries. [3] proposed a probabilistic approach using an objective function derived from information rate distortion theory. MFCC features were used for phoneme segmentation in these studies. Previously we have proposed a phoneme segmentation method using auditory attention (AA) features, which also falls under this category [1]. AA features are biologically inspired and extracted using a model that mimics human auditory attention system. In a sense, the model processes and analyzes the auditory spectrum, and detects the relevant oriented edges and discontinuities in the auditory spectrum corresponding to phone boundaries; i.e. similar to visually observing boundaries in a speech spectrum. In [1], it was shown that the method outperformed the text and model independent methods [2, 3].

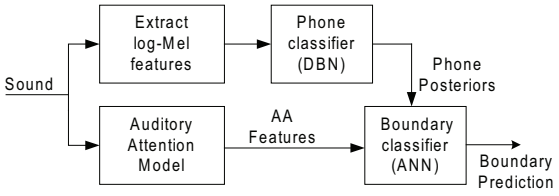

Figure 1: *Fusion of auditory attention (AA) features with phone posteriors (PP) for phoneme segmentation.*

In this paper, to further improve the phoneme segmentation performance, we proposed to combine phone posteriors (PP) with auditory attention features. Phone posteriors are obtained by training a deep belief network (DBN) which estimates phone class posterior scores given acoustic features. It is well known that usually phone classification accuracy drops around boundaries since posterior scores become more confusable with each other; i.e. when it's around a boundary there is no clear winner class, whereas in the middle of a phoneme segment, the winner (i.e. max of posterior scores) is clear cut. This is indeed very useful information for boundary detection purpose and the motivation behind including phone posteriors for phoneme segmentation. It is demonstrated with experiments that combining PP and AA features improves phoneme segmentation performance. The proposed method does not require transcription.

The rest of the paper is organized as follows. The proposed system with AA feature and PP extraction is introduced in Section 2. In Section 3, experiments and results are presented, which are followed by conclusions in Section 4.

2. System Combination for Phoneme Segmentation

The block diagram of proposed automatic phoneme segmentation system is shown in Fig. 1. Auditory attention (AA) features and phone posteriors (PP) are extracted for each audio frame using an auditory attention model and a state-of-the art

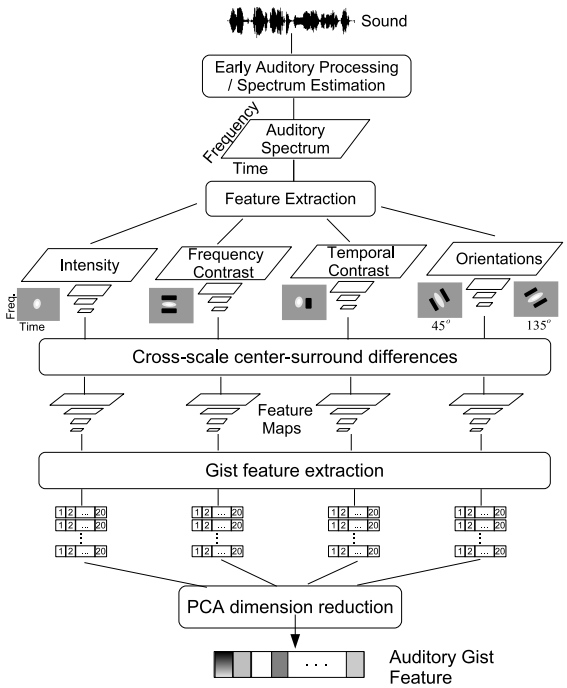

Figure 2: *Auditory Attention Model and Features*

DBN phone classifier, respectively. Then, a boundary classifier, here an artificial neural network (ANN), is trained by fusing AA features with phone posterior scores. During testing, the input sound is similarly processed going through AA feature extraction and PP estimation, and passed through ANN to obtain boundary estimation for that particular frame. Optionally, the output of ANN boundary classifier can be post processed by performing a peak search to obtain segment level results as explained in Section 3. Next, AA feature extraction and PP estimation are described in details.

2.1. Auditory Attention Features

The auditory attention model is biologically inspired and mimics the processing stages in the human auditory system [7, 8]. It is designed to find when and where sound signal attracts human attention. In other words, it captures and detects when there are significant changes in sound characteristics. Hence, it is found to be very effective for change point detection in different tasks including phoneme and syllable segmentation [1, 9], and salient acoustic event detection [10].

The block diagram of the auditory attention model is shown in Fig 2. 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. Next, the auditory spectrum is analyzed by extracting a set of multi-scale features which consist of *intensity* (I), *frequency contrast* (F), *temporal contrast* (T), and *orientation* (O) feature channels. They 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, where the excitation phase and inhibition phase are shown with white and black color, respectively. The RF for I is implemented using a 2D Gaussian kernel, and the RF for F, T, O_θ are 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.

As shown in Fig 2, after extracting features at multiple scales, the "center-surround" differences are calculated resulting in "feature maps". The center-surround operation mimics the properties of local cortical inhibition, and detects the local temporal and spatial discontinuities in feature channels. Center-surround differences are computed using three center scales $c = \{2, 3, 4\}$ and two surround scales $s = c + \delta$ with $\delta\epsilon\{3, 4\}$. Next, an "auditory gist" vector is extracted from the feature maps 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. After augmenting gist vector for each feature map, principal component analysis (PCA) is used to remove redundancy and to reduce the dimension. More details of the model and the parameter choices can be found in [1, 8].

2.2. Estimation of Phone Posteriors with DBN

Phone posterior scores can be estimated using a phoneme recognizer. Recently, the state-of-the art phoneme recognition results on the TIMIT database were achieved using a deep belief network based recognizer [11, 12]. Here, we followed the DBN architecture proposed for phoneme recognition in [12], which combined deep learning with Baum-Welch re-estimation for subphoneme alignment.

25 ms analysis window is used with 10 ms shift to extract log-transformed Mel spectrum coefficients where 26 filter banks are used. Then, each frequency component is normalized to have zero mean and unit variance. To capture context, 21 consecutive frames, which correspond to 210 ms, are augmented resulting in a 546 dimensional acoustic feature.

DBN has one input layer with linear units taking 546 inputs, 3 hidden layers with 1000 binary units, and one output layer with normal logistic units. Original 61 TIMIT phoneme labels are used for training. Each phoneme is modelled assuming 3 temporal states since state representation improves modeling. Then, the output layer of DBN has 61×3=183 outputs. Uniform segmentation is performed at first to obtain state labels, which are realigned later using the Viterbi forced-alignment.

The algorithm is characterized by layers of simple generative models initialized layer-by-layer in an unsupervised way (pre-training), followed by the discriminative retraining of the whole layer using supervised techniques. Pre-training each layer from the lower one to the upper one is done using unsupervised learning algorithm based on the Restricted Boltzmann Machine (RBM), and the final supervised training is done using the well-known error back-propagation algorithm with the standard gradient descent to fine-tune the network for its final classification. Training is repeated until the model parameters converge; in other words until the error in the validation set starts to increase. For more details of DBN training and its parameters one can refer to [12].

Once DBN model is trained, it can be used for extracting phone posterior scores. For that, again log-Mel features are extracted from sound. Then features from 21 frames are augmented for each frame and fed into DBN. At the output of DBN, posterior scores for each state (sub-phoneme) are obtained. To

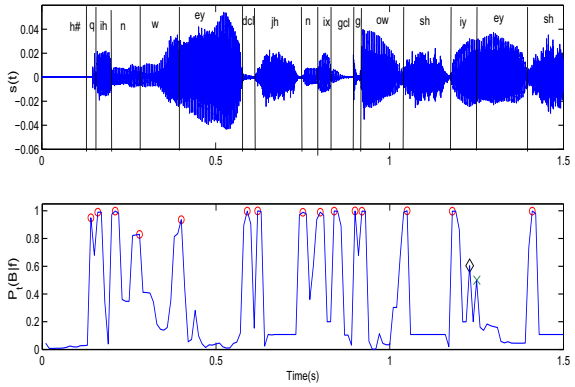

Figure 3: *Phoneme segmentation results for a sample speech segment. "circle", "cross", and "diamond" signs indicate matched, missed, and inserted boundaries, respectively.*

obtain phone posterior scores for each frame, posterior scores of 3 states for each phoneme are simply averaged. For phoneme boundary detection task, phone posterior scores are normalized such that the sum of posteriors for each frame is equal to one.

3. Experiments and Results

TIMIT database is used in automatic phoneme boundary detection experiments since it contains phoneme boundaries that are manually determined by experts. The official train and test splits contain 3696 and 1344 utterances, respectively. The train set is used to train DBN phone classifier and 70.34% frame level phone classification accuracy is achieved on TIMIT official test set.

AA features are estimated every 10 ms using a window of duration W that centers on the current frame to capture the context. In [1], it was found that AA features extracted with $W = 125$ ms performed the best using a grid size of $16 - by - 10$ in the phoneme segmentation task. The length of AA feature was reduced to 77 when 95% of the variance was retained in PCA.

In the boundary detection experiments, a 3-layer ANN is used to learn the mapping between features and phoneme boundaries. ANN has D inputs, $(D + N)/2$ hidden nodes and N output nodes, where D is the length of the feature vector and $N = 2$; i.e. boundary vs. non-boundary. The TIMIT test set is split into two where the core test set is used for evaluating the phoneme segmentation performance and the remaining test set is used for training ANN for boundary classification.

ANN is trained using features for each frame and their corresponding boundary labels; hence the output of ANN returns frame level boundary estimation. However, in the literature, the work on phoneme segmentation has focused on detection at segment/phoneme level rather than frame level. Hence, the following post-processing step is followed to achieve segment level results using the same error margins in [2, 3] for comparison. For each frame, ANN returns a value between $[0, 1]$, which can be considered as $P(B|f)$, the posterior probability of a frame being a phoneme boundary, B, given acoustic features, f. The ANN output score is used to generate 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 have

Table 1: *Phoneme Segmentation Performance for Individual Features Using ANN and Analysis of Context*

Method	Re	Pr	Fs
PP_1fr	39.77	52.83	45.37
PP_3fr	45.51	51.40	48.28
PP_5fr	44.25	52.97	48.22
AA Features	80.59	80.05	81.31

Table 2: *Phoneme Segmentation Performance for Combination of Features Using ANN*

Method	Re	Pr	Fs
AA + PP_1fr	82.52	92.50	87.23
AA + PP_3fr	89.16	87.71	88.43

values larger than a threshold are detected as phoneme boundaries. Figure 3 presents results for a sample speech segment. The first plot displays speech waveform with manually placed phoneme boundaries. The second plot displays $P_t(B|f)$. Here, the threshold was not tuned and simply set to 0.5. As seen in the figure, the method detects all phoneme boundaries within 20 ms of reference boundaries except the boundary between /iy/ and /ey/, since it is misplaced by more than 20 ms. It is also clear from the figure that usually boundaries are detected with sharp peaks and very high probability scores, which indicates that our proposed method is not sensitive to the selected threshold.

For scoring, a time-alignment between the detected phoneme boundaries and the reference ones is used. However, as done in [2], 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. Excessive detected peaks are counted as insertions and having no detected peak for a reference phoneme boundary is counted as a deletion.

First, phoneme segmentation performance of individual features are listed in Table 1 in terms of Recall (Re), Precision (Pr), and F-score (Fs). In the experiments, the effect of context for phone posterior (PP) scores is analyzed by changing the number of frames from one to five (denoted as PP_1fr. and PP_5fr. in the tables) by using neighbouring left and right frames. As shown in Table 1, phoneme segmentation performance using phone posterior scores improves F-score when posterior scores from three frames are augmented (PP_3fr) instead of using scores from a single frame (PP_1fr). Using five context frames did not improve the performance. On the other hand, AA features performed the best achieving 81.31% F-score, whereas PP scores with all context sizes achieved F-score that is under 50%.

Second, phoneme segmentation results achieved by combining AA features with PP for varying context sizes are detailed in Table 2. For all context sizes, phoneme segmentation performance improves significantly when AA features are combined with phone posterior scores. 88.43% F-score is achieved when posterior scores from three frames are combined with AA features (AA + PP_3fr). Although phone posterior scores perform poorly when they are used individually, they improve re-

Table 3: *Comparison of Phoneme Segmentation Methods*

Method	Re	Pr	Fs
Dusan et al [2]	75.2	72.73	73.94
Quiao et al [3]	77.5	78.76	78.13
DBN phone classification	92.7	66.0	77.1
DBN phone recognition	71.8	76.6	74.1
AA Features[1] [1]	80.59	80.05	81.31
AA + PP_3f	**89.16**	**87.71**	**88.43**

sults when they are combined with AA features, indicating that they provide complementary information.

Finally, as shown in Table 3, we compare our results against the ones reported in previous studies [2, 3], which are the state-of-the art to the best of our knowledge. We report two baseline results using boundary estimation from: i) DBN phone classification and ii) DBN phone recognizer. In i), frame level phone classification output is used to predict boundaries; i.e., a boundary exists whenever the phone class changes from a frame to the subsequent frame. In ii), phone posterior scores are passed through a Viterbi decoder for phone recognition, which also outputs estimated phoneme boundaries. As listed in Table 3, DBN frame level phone classifier achieves 77.1% F-score whereas the DBN phone recognizer achieves 74.1% F-score in phoneme segmentation task. We also report AA feature only system as a baseline in Table 3, since it was published originally in [1]. As seen in Table 3, AA feature only system outperforms both [2, 3] and our baselines, DBN phone classifier/recognizer. Finally, the proposed system using AA features and phone posteriors from three frames achieves 88.43% F-score outperforming all other methods including the AA only system. It provides more than 30% relative F-score improvement over the AA only system.

4. Conclusion and Future Work

In this paper, an automatic phoneme segmentation method that combines biologically inspired auditory attention features with phone posteriors is proposed. Phone posteriors are obtained using a state-of-the art DBN based phoneme recognizer. A neural network is used to learn the mapping between phoneme boundaries and combined features. The proposed method detects 89.16% of phoneme boundaries with 87.71% precision achieving 88.43% F-score on TIMIT core test set. It is shown that the proposed method outperforms the recently published text-independent phoneme segmentation methods in [2, 3], baselines created from DBN phone classifier/recognizer, and the system with only AA features in [1]. This indicates that when phone models are available, they provide complementary information to AA features in phoneme segmentation task and improve the performance.

As part of future work, we will investigate using a DBN for phoneme boundary detection instead of a 3-layer ANN to investigate the benefit of using a deeper network. We also plan to conduct experiments in other languages, speaking styles, and noise conditions.

[1] Results with AA are slightly worse than the one in [1], since whole TIMIT training set, which is larger than the amount of data used here, was used for training the ANN in [1].

5. References

[1] O. Kalinli, "Automatic phoneme segmentation using auditory attention features," in *Proc. of Interspeech*, 2012.

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

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

[4] 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.

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

[6] YG Estevan, V. Wan, and O. Scharenborg, "Finding maximum margin segments in speech," in *Proc. of ICASSP*, 2007.

[7] 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.

[8] 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.

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

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

[11] A.-R. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, 2009.

[12] J. Lee and S.-Y. Lee, "Deep learning of speech features for improved phonetic recognition," in *Proc. of Interspeech*, 2011.