Integrating phonetic boundary discrimination explicitly into HMM systems

Yu Wang and Eric Fosler-Lussier

Department of Computer Science and Engineering
The Ohio State University, Columbus, OH, USA
{wangyub,fosler}@cse.ohio-state.edu

Abstract

In this study, we investigate methods of (a) detecting phonetic boundaries directly from acoustics, and (b) integrating these into HMM-based speech recognition. We test the hypothesis that detecting phone boundaries may be easier using phonological features rather than phonetic or direct acoustic information. We also show how HMMs can be more attuned to the transition of phone boundaries by explicitly modeling transition states. Using a 5-state HMM phone model, we improve the accuracy of phone recognition on the TIMIT task.

Index Terms: phonetic boundaries, phonological features, HMM topologies.

1. Introduction

In traditional Hidden Markov Model (HMM) approaches to automatic speech recognition, phonetic states are utilized to model the acoustics of speech. Some of the drawbacks of this modeling are well known: for example, the independence of observations make it difficult to model well both steady-state and dynamic situations in the same kind of model. In essence, phone boundaries are treated the same as the steady-state portions of phones.

However, several approaches have focused particularly on the boundaries between phones. In [1, 2], Morgan *et al.* proposed using Avents (Acoustic events) to model the transitions between phones in a hybrid HMM-ANN system. In this work, neural networks were used to predict phonetic (diphone) boundaries, as well as a non-perceiving state. The authors found that the Avent models were able to represent almost as much information as steady-state representations.

The focus on phonetic boundaries have also been used directly in Gaussian-based HMM systems utilizing *landmark* features. For example, in [3] acoustic observations were allowed to directly affect transition probabilities (instead of just state emission probabilities) in an HMM. By training Gaussian mixtures that focused on the derivative LPC cepstra to condition the probability of transitioning from one phone to another, they were able to improve monophone TIMIT recognition slightly.[1]

The advent of tandem training [4] has conjoined the hybrid and traditional modeling worlds by utilizing neural networks as discriminative front ends for HMM systems. Thus, it is now possible to ask: "Can a discriminatively-trained phone boundary detector be used within traditional HMM methods for improving speech recognition?" In this work, we explore simple methods for incorporating estimates of the presence of phone boundaries into an HTK system [5].

One question that follows from this is what the input to a phonetic boundary detector should be. In recent years, more attention has been given to phonological features in ASR (e.g [6, 7] *inter alia*), such as sonority or place of articulation. Some researchers claim that these features are more flexible in modeling spontaneous speech and more robust in modeling pronunciation variations. If one has a set of phonological feature detectors, the patterns of change in those detectors may lead to good boundary detection.

Motivated by these idea, two experiments are conducted in this paper. The first is to explore various phone boundary detection techniques by comparing the performance of different input features in the task of phone boundary detection (Section 2). After we obtain the boundary information, how to integrate the boundary information into current ASR systems remains a question. We present experiments showing that explicitly modeling the entering and exiting state of a phone as a separate, one frame distribution can improve TIMIT phone recognition, especially when state boundary estimates are included. These results are addressed in Section 3. Conclusions are presented in Section 4.

2. Boundary Detection

2.1. Input features

Three components need to be chosen for a phone boundary detection system: the time-frequency resolution, the extracted features to represent speech signals and the classifi-

[1] The authors mention in this paper is that phone recognition correctness was raised from 58.6% to 62.0% through this method, although an analysis of their results table shows that the corresponding increase in insertions means the accuracy improvement was only from 44.9% to 45.2%.

Attribute	Values
Sonority	Vowel, Obstruent, Sonorant, Syllabic, Silence
Voicing	Voiced, Voiceless, N/A
Manner	Fricative, Stop, Flap, Nasal, Approximant, Nasal Flap, N/A
Place	Labial, Dental, Alveolar,Palatal, Velar, Glottal, Lateral, Rhotic
Height	High, Mid, Low, Lowhigh, Midhigh,N/A
Frontness	Front, Back, Central, Backfront,N/A
Roundness	Round, Non-round, RoundNonround, NonRoundRound
Tense	Tense, Lax, N/A

Table 1: Phonetic attribute classes and their values

cation schemes.

In order to integrate the phone boundary information into existing HMM-based ASR systems, the time-frequency resolution was designed to be a constant, with frames calculated over a 25ms window shifting every 10ms. We assume the presence of labeled training data (either from Viterbi alignment with a pre-trained recognizer, or (as is used here) phonetic labeling provided by linguists). Given the frame definitions and labeling, each frame was labeled as "Left Boundary (LB)", "Right Boundary (RB)" or "Non-Boundary (NB)", depending on its phonetic boundary status. For example, if a sequence of frames have the labels of (/b/, /iy/, /iy/,/iy/), then the phone boundaries are defined as (RB, LB, NB, NB).

Among the choices of input features for boundary detectors, the simplest choice is acoustic features, eg. Mel Frequency Cepstral Coefficients (MFCC) or Perceptual Linear Prediction (PLP) coefficients, which are widely used in ASR systems. 12th order PLP coefficients and the corresponding derivatives were used as a baseline system.

A second choice would be to utilize the tandem methodology [4]: a phone classifier could be trained on the TIMIT training set to predict the phonetic class for every frame. The Multi-layer Perceptron (MLP) trained in this system outputs a 61-dimensional vector for every frame, corresponding to the posterior probability of each monophone. Intuitively, a high distance between any pair of adjacent vectors indicates a phone transition. These vectors could also be used as acoustic features of the subsequently trained boundary detector.

For comparison, we also train phonological feature detectors; Table 1 shows the feature attributes used in this paper, based on articulatory phonology. Given this feature set, each phone can be uniquely defined using eight values, each from every phonetic class. The phone labels of TIMIT were converted into phonetic feature labels; eight feature-extraction MLPs were then trained, one for each phonetic class. The classification results of all trained feature-extraction MLPs were concatenated together to form a 43-dimension phonological feature vector per frame. Similar to the phone-vector situation, one can compare distances between adjacent vectors, or feed the vectors into boundary detector MLPs.

As is noted above, the last component in a boundary detection system is a scheme to detect salient transactions. We compared two methods for estimating boundaries. The first was a simple metric-based method: the Euclidean distance between adjacent frames.[2] A second option was to directly train a pair of MLPs on the binary boundary decision, one for the onset and one for the offset of a phone.

2.2. Experimental setup

All MLPs, including feature-extraction MLPs and phone-boundary-detector MLPs, were trained using ICSI QuickNet MLP software. Each neural network was a three-layer fully-connected MLP. The number of hidden units was chosen according to the number of input units and output units. The whole TIMIT dataset was divided into three partitions. The feature-extraction MLPs were trained on the first partition. Then the trained MLPs are applied to the second partition to generate phonological features; the phone-boundary-detector MLPs were trained on these features. The last partition was used for test. The 61 phones defined in TIMIT were collapsed to 48 phones for training and testing. Since the cross of experimental conditions would generate too many results for presentation here, we only show selected results that demonstrate the detection trends.

2.3. Results

Rather than making a hard decision for each frame, we vary the threshold for detecting boundaries using either the MLP or Euclidean distance method and report the receiver operating characteristic (ROC) curves in Figure 1. Because of the vagaries of our windowing scheme and the subjective decisions of the hand-labeled phonetic transcriptions, when comparing the detected results with hand-labeled phonetic transcriptions, a 10ms tolerance window was used.

[2]We had originally considered using the Kullback-Leibler divergence between vectors since they represent (groups of) probability distributions, but sometimes the probability of some phones/features was zero, causing numerical problems with the KL divergence.

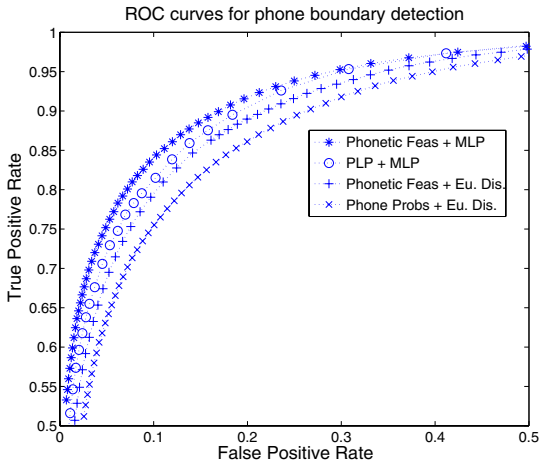

Figure 1: ROC curves of the phone boundary detectors

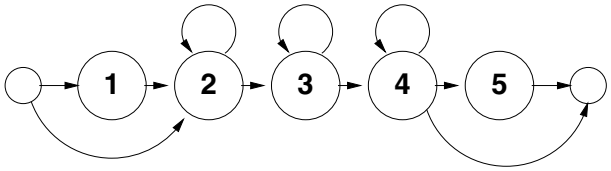

Figure 2: Topology of the 5-state phone model. Small circles indicate non-emitting start/final states.

Three interesting conjectures can be drawn from Figure 1. First, the nonlinear representations learned by the MLP are better for boundary detection than Euclidean-distance metric – phonological features perform better under the MLP regime. Second, using phonological features as an input representation is modestly better than the phone posterior estimates themselves. This is likely because of the redundancy in the estimates along different phonological feature dimensions. Finally, phonological feature representations also seem to edge out direct acoustic representations (PLP).

3. Integrating boundary information into HMMs

If we are to add boundary detection as additional input to the MFCCs in a speech recognizer, the means of the components should be affected by the boundary detection, assuming accurate detection. However, since phone boundaries, as defined in the training of the MLP phone boundary detector, are single-state phenomena, it seems a bit unreasonable to expect a traditional 3-state model to shift its component means significantly. Thus, we introduced a five-state HMM phone model (Figure 2), where the two additional states can catch the phone-boundary transition information, while the 3 self-looped states in the center can model the phone-internal steady regions. The arcs from START to the second state and the arc from the fourth to END are included as escape paths for short phones.

3.1. Experimental setup

3.1.1. Data preparation

The TIMIT dataset was used in this experiment; 39 MFCC coefficients were calculated for each utterance, and binary phone boundaries were extracted by two MLPs, one for the left boundaries and one for the right boundaries, which give us four additional values of boundary information per frame (left_boundary, non_left_boundary, right_boundary, non_right_boundary).

As in the phone boundary detectors introduced in the previous section, the MLPs take 26 PLP coefficients as input vectors and output four boundary-information values. Because of overlap issues with our training sets for boundary detection and TIMIT features, we were unable to use the phonological features + MLP detector; rather, we used the second-best PLP+MLP detector for this experiment. Unlike standard MLPs that use softmax as the activation function applied to the output layer, a linear activation function was applied to the output layer of these MLPs.

Our baseline (system 0) used the standard 39 MFCCs. The first comparison system (1) used 39 MFCCs plus four phone boundary features, which were decorrelated using Karhunen-Loève transformation (KLT). The decorrelation makes it difficult to visualize the effects of the boundary detection on the means, so we also implemented a version without the KLT (system 2). The MFCCs (particularly the velocity and acceleration coefficients) may also be correlated with the binary features, so in system 3 we performed a KLT on the MFCCs and Binary Boundary features jointly. In order to ensure that the KLT was not affecting MFCCs in system 3, we also employed KLT directly on MFCCs (system 4).

3.1.2. HMM training

The HMM training was conducted with the HTK toolkit. Tied-state triphone HMMs were trained on a set of 3696 utterances. A conventional 3-state tied-mixture triphone HMM model was constructed as baseline. The final HMMs had four-Gaussian mixtures.

All four data sets were trained using the same conventional 3-state HMM model. A similar procedures were implemented to train the proposed 5-state model. However, the additional states caused training failure due to data sparsity when reaching the 4-mixture stage. Thus, we adopted a hybrid 2/4 mixture strategy, promoting triphones to 4 mixtures when they had sufficient data to support this.

Similar to the paradigm in [3], the training regime mapped the 61 TIMIT phone set to 48; recognition results

Inputs	3-state ph. accuracy	5-state ph. accuracy
0)MFCC	62.37%	63.41%
1)MFCC+KLT(BinBds)	62.47	63.78
2)MFCC+BinBds	61.25	62.79
3)KLT(MFCC+BinBds)	63.20	64.38
4)KLT(MFCC)	62.70	-

Table 2: TIMIT Phone recognition accuracy

were mapped to the 39-phone set used in [8]. A word-graph grammar was used to enforce triphone constraints.

3.2. Results

Viterbi decoding was used to recognize a test set of 1344 utterances. The baseline system, 39 MFCCs trained on 3-state model achieved accuracy rate of 62.37%. The phone level accuracy rates of all five datasets are illustrated in Table 2.

Several observations can be drawn from the results. After comparing the two columns, we can claim that the proposed 5-state HMM model did perform better than their 3-state counterparts on all training datasets. This is encouraging given that our training of the 5-state model is likely suboptimal because of the hybrid mixture strategy. System 2 allows for interpretation of the state means, and visual inspection of mean of the first binary feature component (not_left_boundary) shows highly negative activation for the state 1 (average: -3.88) compared to states 2 (-2.70) and 3 (-2.33) – evidence that the binary boundary features do often provide accurate information.

For comparison, several experiments were also conducted on a 5-state HMM with a traditional, left-to-right all-self-loops transition matrix. This model achieved accuravy rate of 64.78% on system 1. However, this improvement comes at the cost of vastly increased deletions, showing a bias against short duration phones, whereas the proposed model is balanced between insertions and deletions.

Binary boundaries require the decorrelation to improve recognition (cf. systems 0 and 1 to 2); furthermore, including MFCCs in the decorrelation improves recognition further. The latter system is significantly better for both 3-state and 5-states at $p<0.05$. Comparing the 3-phone system results for systems 0 and 4, it seems that some, but not a lot, of the gain is due to decorrelating the MFCCs.

Overall, the combined strategy of binary boundary features, KLT, and 5-state representations gives almost a 2% absolute improvement in phone recognition. This is rather encouraging since the boundary information we provide is one of the simplest representations; in future work we hope to integrate more refined, multi-class versions of boundary detection, as well as our phonological feature detectors.

4. Conclusion

The work presented here continues the line of argument that phonetic transitions are very important for automatic speech recognition; we have begun exploring the potential space of representations of boundaries. The phone boundary detection experiment shows that phonological features outperform PLP coefficients and phone posteriors as input for detectors. By allowing HMMs to attend to boundaries via the 5-state model, we can also improve recognition when explicit boundary estimates are both present and absent.

5. Acknowledgments

The authors would like to thank Jeremy Morris and Keith Johnson for discussion of this project. This work was funded in part by NSF grant ITR-IIS-0427413; the opinions and conclusions expressed in this work are those of the authors and not of any funding agency.

6. References

[1] N. Morgan, H. Bourlard, S. Greenberg, and H. Hermansky, "Stochastic perceptual auditory-event-based models for speech recognition," in *International Conference on Spoken Language Processing*, Yokohama, Japan, 1994.

[2] S.-L. Wu, "Properties of stochastic perceptual auditory-event-based models for automatic speech recognition," M.S. thesis, UC Berkeley, Berkeley, CA, May 1995, International Computer Science Institute Technical Report TR-95-023.

[3] M. K. Omar, M. Hasegawa-Johnson, and S. Levinson, "Gaussian mixture models of phonetic boundaries for speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Trento, 2001.

[4] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature stream extraction for conventional HMM systems," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Beijing, 2000.

[5] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *HTK: Hidden Markov Model Toolkit, v3.3*, Cambridge University Engineering Department, Cambridge, UK, 2005, Available at http://htk.eng.cam.ac.uk.

[6] M. Hasegawa-Johnson et. al, "Landmark-based speech recognition: Report of the 2004 Johns Hopkins Summer Workshop," in *International Conference on Acoustics Speech and Signal Processing*, Philadelphia, 2005.

[7] M. Rajamanohar and E. Fosler-Lussier, "An evaluation of hierarchical articulatory feature detectors," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, San Juan, Puerto Rico, 2005.

[8] K. Lee and H. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.