Novel Entropy based moving average refiners for HMM Landmarks

*Rahul Chitturi, Mark Hasegawa Johnson**

University of Texas at Dallas, USA
rahul.ch@crss.utdallas.edu

* University of Illinois at Urbana Champaign, USA
jhasegaw@uiuc.edu

Abstract

The training of precise speech recognition models depends on accurate segmentation of the phonemes in a training corpus. Segmentation is typically performed using HMMs, but recent speech recognition work suggests that the transient acoustic features characteristic of manner-class phoneme boundaries (landmarks) may be more precisely localized using acoustic classifiers specifically designed for the task of landmark detection. This paper makes an empirical exploration of entropy based moving average techniques that are capable of improving the time alignment of phoneme boundaries proposed by an HMM-based speech recognizer. On a standard benchmark data set (A database of Hindi – National Language of India), we achieve new state-of-the-art performance, reducing RMS phone boundary alignment error from 28ms to 15ms.

Index Terms: Entropy, Moving Average, Landmark, Segmentation

1 Introduction

The automatic segmentation of speech especially in real world noisy environments is a challenging problem. Most importantly, the efficiency achieved in automatic detection of speech boundaries largely determines the accuracy of the recognition and synthesis engines. Even minor improvement in speech boundary detection front-end greatly influences the overall system accuracy in the long run. Accurately segmented training data improves the precision of both speech recognition and speech synthesis models. Phonetic segmentation has also been proposed as part of a lattice rescoring algorithm for multipass speech recognition [1]. Since manual segmentation of speech is time consuming and unrealistic in most conditions, various approaches on automatic speech segmentation have been proposed [2, 3], most typically including forced alignment of an HMM-based sentence model [4, 5, 6] observing standard speech recognition features such as energy, LPCC, MFCC, and PLP.

Testing perceptual quality of speech synthesis and the recognition accuracy is beyond the scope of the present paper (because of space constraints); instead this paper will focus on improvements in speech segmentation accuracy.

The most commonly used method of endpoint detection is the use of short-time or spectral energy [7, 8]. A newer promising approach involves the use of entropy to find endpoints. The main features of an entropy profile include less sensitivity to the changes in the amplitude of the speech signal, which directly results in retention of more detail as compared to the corresponding energy profiles, thus taking care of noise.

This article proposes the use of heterogeneous classifiers, in a second-pass speech segmentation system. First segmentation is done using HMM techniques and then the segmented data is sent to the phone boundary refiner which essentially used entropy based moving average techniques for refinement of phone boundaries.

2 Related Work

Maximum Entropy principle was first applied to language modeling in [9]. Later on speech/music segmentation posterior probability based features introduced in (Williams and Ellis, 1999), namely entropy and dynamism which were explored in 2003 by Ajmera et al. [10]. Recently in 2005[11], Maximum entropy was used in Topic segmentation and word level utterance segmentation. This work further explores the application of Maximum entropy for segmentation at phone level.

Generally the moving averages are used in the stock market domain to identify the sudden changes in the prices of the stocks. The advantages of using this method are that it is very effective in catching the sudden changes and the threshold setting is dynamic. The threshold depends on the context of the previous frames. So, the common problems with the threshold setting like specific threshold for specific conditions (in our case for different phones) will be taken care by this method.

3 Database Description

All our experiments were conducted on the Hindi (National Indian Language) database, which has over 600 sentences collected from a native speaker and contains substantial coarticulatory effects. This database was collected at a 16 KHz sampling frequency, 256 Kbps bit rate and a single channel, using a headset. This database has well defined transcriptions corresponding to the speech data and even the phone-level alignment was done manually. For the purpose of exact evaluation of our algorithms, we used the manual phone-alignments, but we don't assume that these are necessary.

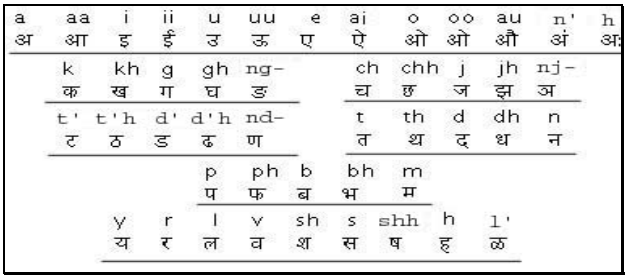

Figure 1- Phones and their IT3 Notation

4 Baseline System

The goal of the algorithms proposed in this paper is to refine the phoneme boundary times proposed by the HMM, in order to reduce their RMS error with reference to a "ground truth" manual phonetic transcription. Though HMMs are known to work better for tri-phone models, since we are concerned only with the time boundaries of a particular phone, monophone based HMM segmentation system is done. The 39-element feature vector includes 12 MFCCs plus sum-squared signal energy, with delta and delta-delta coefficients appended. There are a total of 48 phones as shown in Fig. 1. Each phone is modeled by 3 emitting states, with 3 Gaussians per state. HMMs were initialized flat (to the same mean and variance), then HMMs were trained using embedded re-estimation for 15 iterations. Finally, forced alignment was done to get the phone boundaries. This system achieved an RMS phone boundary alignment error of 28.8ms

5 Landmark Refinement Techniques

5.1 Refinement Using Moving Averages

"Change detection" is one of the most routine tasks in market and financial analysis; millions of dollars can be made by rapidly detecting significant changes in the behavior of a stock price or market index. The most ubiquitous method of change detection is to compare the instantaneous value of a function with its short-term and long-term moving averages. The problem of speech segmentation is also a problem of "change detection," because there is typically a lot of entropy change at a phone boundary, especially at the landmark separating phones of different manner classes. Fig. 2, for example, shows the entropy variations in a sample utterance.

If s[n] is entropy, the moving average is calculated as follows:

$$mov_{avg}[i] = \frac{1}{N}\sum_{k=i-N}^{i}(s[k]) \qquad (1)$$

where N is the window length. Since the moving average is delayed by N/2 samples, sudden movements in the entropy of the speech signal cause the entropy to diverge from its moving average [12] as shown in Fig. 2. Landmarks can therefore be detected as follows:

$$Signal[i] = \begin{cases} 1 & if\ (|s[i]-mov\_avg[i]|) > \textbf{\textit{trsh}} \\ 0 & otherwise \end{cases}$$

where ***trsh*** is a very small positive threshold value taken as $mov\_avg[i]/100$ in our case. The threshold is dynamic as depends on the context of previous frames. Generally the thresholds will be different for different phones, but using this method we can set a global threshold for all the phones because this is dependent on the previous frames. This is the main advantage of using this method. After doing this analysis over the entire speech signal, we implement the following algorithm (Procedure (1)) to refine the boundaries.

```
for all HMM phone boundaries l_i in the signal
    for all the frames f_ij of 5 ms in the
    window of ( l_i – 40) ms to (l_i+20) ms
        if signal[f_ij] = = 1
            l_i = f_ij
            goto l_i+1
        endif
    end
end
```

Procedure 1

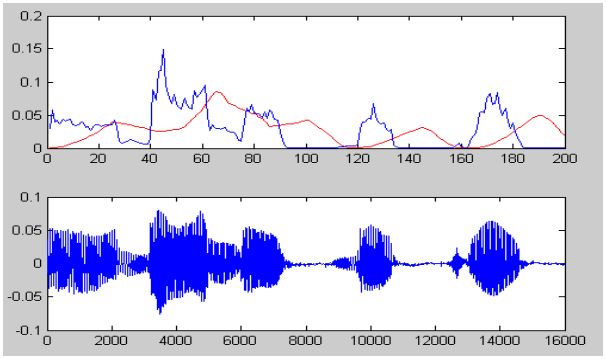

Fig.2. Entropy compared to its moving average.

5.2 Formant Analysis + Moving Average

Formants (resonant frequencies of the vocal tract) are a useful acoustic parameter because they have the flexibility to model both linguistic and extra-linguistic processes [6]. Formant estimation has been proposed for speech enhancement in noisy environments [13]. Previous landmark-based speech recognition systems have shown that SVMs observing MFCCs+formants tend to dramatically outperform SVMs observing MFCCs alone [1].

In this paper, formants are approximated locally as the roots of the LPC polynomial. Manner-change landmarks cause rapid changes in the roots of the LPC polynomial, therefore we can refine the landmarks by using the following procedure (2).

```
for all HMM phoneme boundaries l_i in the signal
    for all the frames f_ij of 5 ms in the
    window of ( l_i – 40) ms to (l_i+20) ms
        if (prev_f1 - f1)>1000 or
          (prev_f2 - f2)>1000 or
          (prev_f3 - f3)>1000
            l_i = f_ij
            goto l_i+1
        endif
    end
end
```

Procedure 2: Notation same as Procedure 1

where *prev_fn* = frequency of the nth formant in the previous frame, and *fn* = frequency of the nth formant in the present frame.

Method	Average. Deviation(ms)	<5 ms	<10 ms	<15 ms	<20 ms
HMM	28.8	3.82%	8.70%	18.06%	34.35%
MA	19.77	33.05%	49.66%	59.60%	67.31%
Formant + MA	19.01	29.40%	50.12%	62.19%	71.41%
Entropy	17.27	28.90%	52.68%	67.38%	77.03%
Entropy + MA	15.44	40.46%	61.20%	72.49%	79.21%

Table 1: Accuracy of phone boundary refinement algorithms

Category	< 5ms	< 10ms	< 15ms	< 20ms
Obstruent-Obstruent	16.54%	32.59%	43.2%	53%
Obstruent-Nasal	41.93%	64.51%	74.19%	77.42%
Obstruent-Vowel	56.71%	79.17%	86.98%	90.65%
Obstruent-Semivowel	41.88%	59.88%	74.06%	82.3%
Nasal-Obstruent	24.18%	44.88%	60.69%	69%
Nasal-Nasal	17.85%	26.78%	33.92%	37.5%
Nasal-Vowel	35%	56.4%	69.25%	77.22%
Nasal-Semivowel	23.21%	32.14%	48.21%	64.28%
Vowel-Obstruent	48.56%	75.02%	86.53%	92.04%
Vowel-Nasal	35.55%	57.45%	70.87%	80.27%
Vowel-Vowel	8.059%	15.52%	21.19%	29.25%
Vowel-Semivowel	33.93%	57.27%	72%	79.59%
Semivowel-Obstruent	38.63%	65%	77.27%	81.8%
Semivowel-Nasal	23.07%	40.38%	57.59%	71.15%
Semivowel-Vowel	23.16%	41.54%	58.16%	68.9%
Semivowel-Semivowel	28.39%	44.44%	56.09%	60.49%

Table 2: Accuracy of the Entropy+MA phone boundary refinement algorithm, as a function of the manner classes of phones on either side of the boundary

5.3 Entropy Based Analysis

Entropy is a measure of the randomness or instability in a system. A phone boundary, and especially a manner-change landmark, can be defined as an instant of acoustic instability. Therefore, the entropy of the sample within a 7-frame window is a good measure of the probability that a landmark is present. In this paper we use a Gaussian probability density function for the entropy of the speech signal (s[n]), thus:

$$\frac{1}{\sigma'\sqrt{2\pi}}\exp\left(-\frac{(s[x]-\mu')^2}{2\sigma'^2}\right) \qquad (4)$$

where $\sigma' = \sigma(s[n-N..n])$ which is the variance of short term energies of a speech signal in a window of length N , $\mu' = \mu(s[n-N..n])$ which is the mean of short term energies of a speech signal in a window of length N and $x \in \{n-N..N\}$ and the entropy of the sample is given by $\ln\left(\sigma'\sqrt{2\pi}\right)$. Then running the following procedure (3) refines the boundaries using entropy.

for all HMM phoneme boundaries $\boldsymbol{l_i}$ *in the signal*
In all the frames f_{ij} *of 20 ms in the*
window of ($\boldsymbol{l_i}$ *– 40) ms to (*$\boldsymbol{l_i}$*+20) ms*
estimate the frame $\boldsymbol{f_{ik}}$ *which has the maximum entropy.*
$\boldsymbol{l_i} = \boldsymbol{f_{ik}}$
end

Procedure 3: Notation same as Procedure 1

5.4 Entropy Analysis + Moving Average

Our best results were achieved by using entropy to correct systematic errors in the HMM output, then using the moving average to correct any residual non-systematic errors. Systematic errors may exist because of inaccuracies in the HMM Gaussian observation PDFs (e.g., so that the initial state in each phone accepts spectra from neighboring phones), because of peculiarities in the alignment of HMM time stamps with the centers of their corresponding cepstral or delta-cepstral computation windows, because of the tendency for the spectrum of a speech frame to look like the spectrum of the highest-energy samples in that frame, or for other reasons beyond the immediate control of the experimenter. We correct for systematic error by finding the offset between every HMM phone boundary and its nearby high-entropy frame, by averaging these offsets over the entire speech corpus, and then

by shifting each HMM phone boundary time by the corpus-average maximum-entropy offset. After correcting for systematic error in this way, the method of Sec 5.1 is applied.

6 Algorithm Comparison

This section compares the RMS error in milliseconds of the algorithms described in the Sec. 5. Results are shown in Table 1. The percentage of phones whose error is less than 5ms, 10ms, 15ms and 20ms are shown in the respective columns. The entropy-based method was the most accurate one-step phone boundary refinement algorithm. Best total performance was achieved by the entropy + MA algorithm described in Sec. 5.4. Since the results vary with the speech database, and as there is no benchmark speech database for the Hindi Language, we have only compared our algorithms with the methods that are usually employed like HMMs, on the database that was described in Section 3. The train data set has nearly 500 sentences and the test data set has 100 sentences, each sentence having approximately 40-50 phonemes (10-15 words).

The phonetic performance of the entropy + moving average technique is shown in Table 2 as a function of the manner class (obstruent consonant, nasal, semivowel, or vowel) of the phones on either side of the boundary. As expected, acoustic phonetic landmarks (defined as boundaries between phones of differing manner class) are universally detected with better precision than are boundaries between phones of the same manner class. Nevertheless, the proposed method is even able to detect boundaries between phones of the same manner class with a modal error of 20ms

7 References

[1] M. Hasegawa-Johnson, J. Baker, S. Borys, K. Chen, E. Coogan, S. Greenberg, A. Juneja, K. Kirchhoff, K. Livescu, J. Muller, K. Sonmez, and T. Wang, *Landmark-Based Speech Recognition: Report of the 2004 Johns Hopkins Summer Workshop.* Technical report .

[2] Y.-J. Kim, A. Conkie, “Automatic Segmentation Combining an HMM-Based Approach and Spectral Boundary Correction,” in: Proc. ICSLP2002, Denver, Colorado, Session TuA4p.13, 145- 148, 16-20 Sept. 2002.

[3] Syrdal, AK, Hirschberg, J., McGory, J. and Beckman, M., "Automatic ToBI prediction and alignment to speed manual labeling of prosody," In Speech. Communication, vol. 33, 135-151.

[4] K. Tokuda, H. Zen, A.W. Black, ``An HMM-based speech synthesis system applied to English," *IEEE Speech Synthesis Workshop*, 2002.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, ``Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," *EUROSPEECH*, pp.2347-2350, 1999

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

[7] Junqua J. C., Mak B., and Reaves B., "A Robust Algorithm for Word Boundary Detection in the Presence of Noise", IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 3, Apr. 1994. pp. 406-412

[8] Ganapathiraju A., Webster L., Trimble J., J. Bush J. and J. Kornman J., "Comparison of Energy-Based Endpoint Detectors for Speech Signal Processing," *Proceedings of the IEEE Southeastcon*, Tampa, Florida, USA, April 1996.pp. 500-503

[9] A. Berger, S. Della Pietra and V. Della Pietra, “A maximum entropy approach to natural language processing”, *Computational Linguistics*, vol. 22, no. 1, pp.39-71, 1996.

[10]Speech/Music Discrimination using Entropy and Dynamism Features in a HMM Classification Framework, *J. Ajmera, I. McCowan, and H. Bourlard*, in "Speech Communication", 2003.

[11] H. Christensen, B. Kolluru, Y. Gotoh, and S. Renals. Maximum entropy segmentation of broadcast news. In *Proc. IEEE ICASSP*, 2005.

[12] J.F. Kenney and E.S. Keeping, "Moving Averages." §14.2 in *Mathematics of Statistics, Pt. 1, 3rd ed.* Princeton, NJ: Van zostrand, pp. 221-223, 1962.

[13] M. Tang, *Large Vocabulary Continuous Speech Recognition Using Linguistic Features and Constraints.* Ph.D. Thesis, MIT Department of Electrical Engineering and Computer Science, May 2005.