On the Relation between Maximum Spectral Transition Positions and Phone Boundaries

Sorin Dusan and Lawrence Rabiner

Center for Advanced Information Processing
Rutgers University, Piscataway, New Jersey, U.S.A.
{sdusan, lrr}@caip.rutgers.edu

Abstract

Earlier research has shown that the maximum spectral transition positions are related with the perceptual critical points that contain the most important information for consonant and syllable perception. This paper presents a quantitative analysis of the relation, in time, between the maximum spectral transition positions and the phone boundaries in fluent read speech. This analysis is based on the training part of the TIMIT American English database which contains both phone boundaries and labels manually-determined by a group of experts. The results of this analysis show that there is a significant correlation between the maximum spectral transition positions and the manually selected phone boundaries. This, in turn, suggests that there is an important relation between the commonly accepted phone boundaries and the perceptual critical points.

Index Terms: phone boundaries, spectral transition, phonemes

1. Introduction

Accurate phone boundaries are important (and essential) for acoustic-phonetic analysis, automatic speech recognition (ASR), and speech synthesis systems. However, the process of manually determining phonetic transcriptions and segmentations is laborious, expensive, and requires expert knowledge. In addition, there is always some disagreement among the experts with respect to the exact position, in time, of some phone boundaries. Because the cost and effort required for this process are significant for large databases, the need to automate it is mainly motivated by the need for large speech databases used to train and evaluate new ASR systems or to build concatenative text-to-speech (TTS) systems. Many automatic phonetic segmentation and labeling methods have been proposed for such purposes. References [1] and [2] provide overviews and comparisons of such methods, most of which are based on forced recognition and alignment starting from the orthographic transcription of the speech material.

Automatic segmentation and detection of the phone boundaries within a sentence can also be achieved by employing simpler algorithms based on acoustic rate of change or boundary models [3], [4]. Such methods do not involve a full phonetic recognition process and they do not provide the detailed phonetic transcription of the sentence.

Regardless of the type of method used for automatic phonetic segmentation, evaluation of the accuracy of these methods usually relies on a manually segmented and labeled speech database. However, the criteria and cues used in the automatic and manual phonetic segmentation, respectively, need not be exactly the same. For example in the utterance "she'' a labeler might place a boundary at the beginning of voicing whereas an automatic boundary detector, based on spectral rate of change, might place the same boundary at a position which corresponds to the peak in the spectral rate of change. Previous work has shown that the maximum spectral transition positions are in close proximity (within approx. 10 ms) to the perceptual critical points that carry the most important information for consonant and syllable perception [5]. This paper presents a quantitative analysis of the relation between the maximum spectral transition positions and the manually obtained phone boundaries in fluent read speech.

2. Method

This quantitative analysis is achieved by comparing the phone boundary positions obtained from a manually segmented speech database with the boundary positions obtained by means of an automatic segmentation method based on maximum spectral rate of change. It should be noted that the resolution of the automatic method depends on the frame step (10 ms in this paper) of the analysis window, whereas the resolution of the manual method is at the audio sampling step (0.0625 ms).

2.1. Speech corpus

The analysis performed in this study was done using the training part of the TIMIT American English acoustic-phonetic corpus [6]. This database contains utterances from 462 speakers, each reading 10 sentences. The transcription uses 61 phonetic symbols for segmentation and labeling. Not all of these 61 symbols represent phonemes in American English; e.g., the stop consonants are represented as two separate segments and symbols: one for stop closure and one for stop burst. This dataset contains 172,460 between-phone boundaries manually determined by experts. These boundaries do not include the boundaries placed at the beginning and end of the sentences.

2.2. Spectral features

The spectral features used in this study are the Mel-Frequency Cepstrum Coefficients (MFCC). These spectral features are extensively used in ASR and details on their computation can be found in [7]. For each sentence in the database, the speech signals are first transformed into spectral frames (computed over 32 ms Hamming windows) and then transformed into a set of 10 MFCC coefficients (excluding the zero order coefficient that represent the total energy). The total

energy coefficient was not used here because this analysis focuses on the spectral features alone. The frame rate employed in this study was 100/s (10 ms frame step or frame increment). Also the dynamic MFCC coefficients were not directly used in this study since the spectral rate of change represents a dynamic measure by itself.

2.3. Spectral transition measure

The criterion used in this study for phonetic segmentation was based on a measure of the spectral rate of change in time. Since the spectral rate of change usually displays peaks at the transition between phones, such a measure can be used to detect boundaries between phones. It should be noted that not all the peaks in the spectral rate of change correspond to a valid boundary between phones. For example, a diphthong would display a peak of the spectral transition measure approximately located at its center; however this does not represent a valid phone boundary.

The spectral transition measure employed in this study was the same as that proposed in [5] and it can be interpreted as the magnitude of the spectral rate of change. This spectral transition measure (STM), at frame *m,* can be computed as a mean-squared value

$$STM(m) = (\sum_{i=1}^{D} a_i^2(m)) / D, \quad (1)$$

where D is the dimension of the spectral feature vector (10 in this case) and $a_i(m)$ is the regression coefficient or the rate of change of the spectral feature $MFCC_i$ defined as

$$a_i(m) = (\sum_{n=-I}^{I} MFCC_i(n+m) * n) / (\sum_{n=-I}^{I} n^2), \quad (2)$$

where n represents the frame index and I represents the number of frames (on each side of the current frame) used to compute these regression coefficients. We use $I=2$ for a 10 ms frame step corresponding to an interval of 40 ms centered on the current frame at which the STM value is computed. A larger interval could result in missing some phone boundaries whereas a shorter interval could result in the detection of too many false phone boundaries.

2.4. Boundary detection

The detection of the phone boundaries used here involves two steps: a peak picking method and a post-processing method for removing spurious (false) boundaries. First, all the peaks in the spectral transition measure, computed every frame, are marked as possible phone boundaries. Then the boundaries corresponding to the peaks that are not higher than the adjacent STM values by at least 1% of the highest peak in each sentence are removed. The 1% threshold was determined experimentally. A second criterion for removal of spurious boundaries is to compare the STM peak values with those of the adjacent valleys on both sides. The valleys usually occur at much larger distance than the adjacent frames used in the first part of the post-processing. If the difference between the value of each peak and the values of its adjacent valleys (on both sides) is not larger than 10% of the peak value, then that peak corresponds to a flat STM region and it is removed from the boundary list. Each of the automatically detected phone boundaries is placed in time at a frame position (multiple of the frame step). No attempt was made to remove the spurious phone boundaries detected in the central region of diphthongs and similar sounds.

3. Analysis results

In order to perform the comparison between the manually and automatically placed phone boundaries the former are converted to the closest adjacent frame positions. This is done because the frame sampling is at multiples of the frame step size and the manual boundaries are distributed uniformly within the frame step intervals. Thus this boundary conversion to the closest frame positions induces an average absolute difference equal to a quarter of the frame step and a maximum absolute difference equal to half of the frame step. Because there is no way to eliminate this fixed difference, for a given frame step, the comparison of the automatically detected boundaries is done with the converted boundaries and not with the original boundaries.

Figure 1 presents typical results of the automatic phone boundary detection for the first 1 s portion of a TIMIT sentence for a frame step size of 10 ms. The first plot at the top displays the speech signal and the manually placed phone boundaries (vertical bars) and labels. The abscissa represents the time in sec. The second plot displays the STM values for the first 100 frames and the automatically detected phone boundaries after post-processing. The third plot displays the STM values and the missed phone boundaries. The fourth plot displays the STM values and the inserted (spurious) phone boundaries. The abscissa in each of the last three plots is the frame index. In the third plot one can see that the detection algorithm missed three boundaries. These correspond to three phones whose boundaries are difficult to detect using this method: stop burst /d/, stop closure /kcl/, and flap /dx/. An STM peak is totally missing between /d/ and /ow/, between /kcl/ and /m/, and between /dx/ and /ix/ whereas the peak between /ih/and /dx/ is very small and flat and it was removed from the phone boundary list.

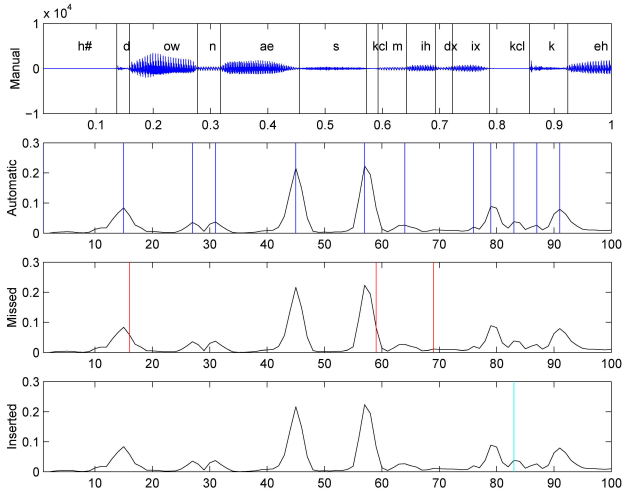

Figure 1 R*esults of the automatic detection process.*

In order to observe the systematic behavior of the automatic detection method, the results of processing another token of the same sentence but from a different speaker are presented in Figure 2.

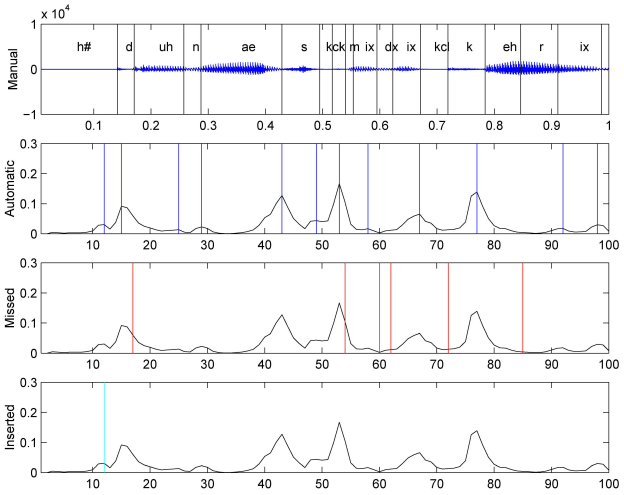

Figure 2 *Another example of results for the automatic detection process.*

One can see in both figures that the most accurate phone boundaries correspond to the fricative /s/, which is delimited by two strong STM peaks. In both figures the boundary between the stop burst /d/ and the following vowel is missing, as well as the boundary delimiting the end of /kcl/ and the boundaries of the flap /dx/. These boundaries correspond in general to very short (transient) speech events. If the detection algorithm increases its time resolution by decreasing the frame step and the STM computation interval ([-I,I] from Eq. 2) then spurious boundaries will be inserted due to more "noisy" STM values. Such spurious boundaries are frequently detected in the /h#/ and /sil/ segments (an example is shown in Figure 2).

Table 1 presents boundary counts and percentages for the automatic phone boundary detection experiment. The manually derived boundaries are represented as (Man.) and the automatically derived boundaries are represented as (Aut.).

Table 1. *Automatic phone boundary detection results*

	Total (Man.)	Detected (Aut.)	Missed (Aut.)	Inserted (Aut.)
Count	172,460	145,950	26,510	48,566
Percent	100%	84.6%	15.4%	28.2%

Approximately 85% of the manually placed phone boundaries from the training part of TIMIT were detected by the automatic method based on the spectral transition measure alone. Thus, about 15% of the original manual boundaries were missed, but another 28% spurious boundaries were inserted. By removing all the missed and inserted boundaries a more detailed analysis can be made by examining the deviation in time between the detected boundaries at the maximum spectral transition positions and the manually placed boundaries.

Figure 3 presents a normalized histogram of the absolute deviations between the 145,950 automatically detected boundaries and the corresponding 145,950 manually placed boundaries (26,510 boundaries were removed).

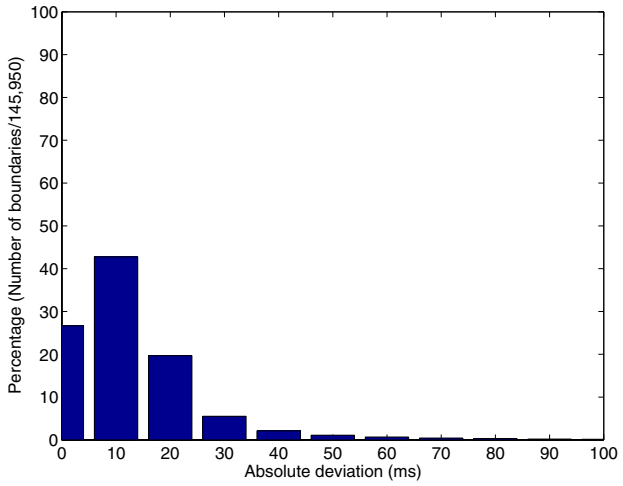

Figure 3 *Normalized histogram showing the absolute deviation between the145,950 automatically detected boundaries and the corresponding 145,950 manually placed boundaries.*

It can be seen that about 27% of the automatically detected boundaries (using the maximum spectral transition criterion) coincide with the manually placed boundaries. Moreover another 43% are within 10 ms of the manual boundaries and another 20% are within 20 ms. Figure 4 presents a normalized cumulative histogram derived from the histogram in Figure 3.

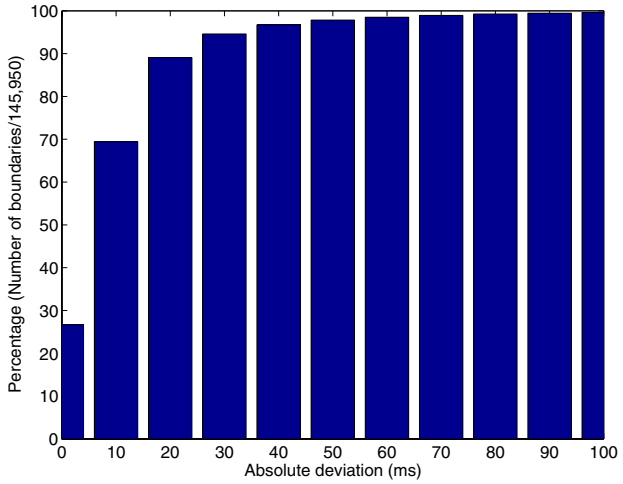

Figure 4 *Normalized cumulative histogram showing the absolute deviation between the145,950 automatically detected boundaries and the corresponding 145,950 manually placed boundaries.*

From this last figure it can be seen that about 70% of the automatically detected boundaries are within 10 ms of the manually placed boundaries, 89% are within 20 ms, 95% are within 30 ms, and 97% are within 40 ms, respectively.

4. Discussion

The above results suggest that there is a significant correlation between the phone boundaries determined by experts and the

corresponding boundaries automatically detected using the maximum spectral transition criterion. These results are based on data from all the 4,620 sentences from the 462 speakers and over all types of phones. A similar analysis was performed on the same database but with a frame step of 5 ms. The overall results were better for the 10 ms frame step. Hence we will not discuss results using a 5 ms frame step in this paper.

A brief comparison with earlier results shows the following. In [3] the percentage of detected boundaries within 20 ms from the manually placed boundaries (using the same TIMIT database) was between 60.5% and 96.1%, depending on the type of transition, with an overall average of 73.8%. This performance is significantly below the 89% figure obtained here for concurrence in boundary to within 20 ms.

In reference [8] results on automatic alignment of phonemic labels to within an interval of 20 ms were 88.1% when tested on the data used for training and 82.3% for unseen data. The detection methods in both [3] and [8] used training data to train the detection models whereas the method used in this paper does not use phonetic or transitional models for boundary detection.

Other results on TIMIT using automatic aligners have been reported in [9] and [10]. In [10] it was shown that 71% of the aligned boundaries using the *Aligner* [9] were within 16 ms, 90% were within 32 ms, and 97% were within 64 ms from the manually placed boundaries. The results in [1] using a standard HMM-based segmentation algorithm show 85.9% of the detected boundaries to be within 20 ms from the manually derived boundaries. As discussed above, the comparable results from the automatic detection method described in this paper are significantly better. One mitigating factor in [10] was that the speech was sampled at an 8 kHz rate rather than at the 16 kHz rate used in this paper. However, there might be other differences among these methods with respect to how the boundaries were counted and how the models were trained.

5. Conclusions

This paper presented a quantitative analysis of the relation between the phone boundaries automatically detected at the maximum spectral transition positions and the manually detected boundaries in the training part of the TIMIT database. This analysis represents an initial approach to a more refined analysis and to a development of an automatic phonetic boundary detector. Such phonetic boundary detector is intended to be part of a larger research project focusing on non-conventional ASR [11]. As yet there have been very few measures taken to remove the spurious (false) boundaries and no measure taken to reduce the missed boundaries while maintaining a small number of inserted boundaries. Future work will focus on these problems and may employ additional acoustic features.

Another future direction is to perform a more detailed analysis of the phonetic segments and labels that have a high rate of missed or inserted boundaries. A similar analysis can be performed to see the degree of deviation of the detected boundaries from the manually placed boundaries with respect to particular types of phonetic transitions (e.g. stop burst to vowel, fricative to nasal, etc.).

Since it was shown in [5] that there is a close proximity between the maximum spectral transition positions and the perceptual critical points that carry the most important information for consonant and syllable perception, it appears that, based on the current results, there is also a close relation between the commonly accepted phone boundaries and the perceptual critical points. Complementary to the results from [5] it was recently shown in [12] that indeed the manually placed phone boundaries contain very significant information about the vowel identity, especially in the dynamic spectral features. However, more studies are required in order to fully clarify this.

6. Acknowledgements

This work was supported in part by an ITR grant (IIS-04-27413) from the U.S. National Science Foundation.

7. References

[1] Cucchiarini, C. and Strik, H., "Automatic Phonetic Transcription: An Overview," Proc. of Int. Congress of Phonetic Sciences, pp. 347-350, 2003.

[2] Demuynck, K. and Laureys, T., "A Comparison of Different Approaches to Automatic Speech Segmentation," Proc. of the 5th Int. Conf. on Text, Speech and Dialogue, pp. 277-284, 2002.

[3] Micallef, P. and Chilton, T., "Automatic Identification of Phoneme Boundaries Using a Mixed Parameter Model," Proc. EUROSPEECH'97, Rhodes, Greece, 1997.

[4] van Santen, J. P. H. and Sproat, R. W., "High-Accuracy Automatic Segmentation," Proc. of EUROSPEECH'99, Budapest, Hungary, 1999.

[5] Furui, S., "On the Role of Spectral Transition for Speech Perception," J. Acoust. Soc. Amer., Vol. 80(4):1016-1025, 1986.

[6] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L., "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus," U.S. Dept. of Commerce, NIST, Gaithersburg, MD, 1993.

[7] Davis, S. and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition," IEEE Trans. on Acoust., Speech, and Signal Process. 28 (4), 357-366, 1980.

[8] Svendsen, T. and Kvale, K., "Automatic Alignment of Phonemic Labels with Continuous Speech," Proc. of ICSLP'90, Kobe, Japan, pp. 997-1000, 1990.

[9] Wightman, C. and Talkin, D., "The Aligner: A system for Automatic Alignment of English Text and Speech," Software and Documentation, Entropic Research Laboratory, Washington DC, U.S.A., 1994.

[10] Pellom, B. L. and Hansen, J. H. L., "Automatic Segmentation of Speech Recorded in Unknown Noisy Channel Characteristics," Speech Communication, 25, pp.97-116, 1998.

[11] Hou, J., Rabiner, L., and Dusan, S., "Automatic Speech Attribute Transcription (ASAT) – The Front End Processor," In Proc. of ICASSP'06, Toulouse, France, 2006.

[12] Dusan, S., "On the Distribution of Information and Intrinsic Variability for Classification of Coarticulated Vowels," In Proc. of ITRW on Speech Recognition and Intrinsic Variation, Toulouse, France, 2006.