On Designing and Evaluating Speech Event Detectors

Jinyu Li and *Chin-Hui Lee*

School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA, USA
{jinyuli, chl}@ece.gatech.edu

Abstract

We study issues related to designing speech event detectors for automatic speech recognition. Event detection is a critical component of a recently proposed automatic speech attribute transcription (ASAT) paradigm for speech research. Similar to keyword spotting and non-keyword rejection, a good detector needs to effectively detect speech attributes of interest while rejecting extraneous events. We compare frame and segment based detectors, study their properties in detecting manners of articulation, and propose new performance measures. We test these detectors on the TIMIT database with several evaluation criteria. Our results indicate that segment based detectors outperform frame based detectors in several key aspects of speech detector design. We also show that the performance can be significantly enhanced by incorporating discriminative training into designing speech event detectors.

1. Introduction

A detection-based automatic speech recognition (ASR) paradigm through automatic speech attribute transcription (ASAT) has recently been proposed [1]. In this framework, a bank of speech event detectors that are capable of producing consistent detection results need to be developed. These "events" are usually low level speech attributes related to information required to form higher level "evidences", so they can then be combined to detect phones, words and sentences, and perform speech recognition in a probabilistic manner.

In this study, a speech event is defined as the presence of a particular acoustic-phonetic attribute. The task of detecting speech events is usually more difficult than conventional signal detection, in which the target signals can often be clearly characterized. Classical signal detection theory has thus been well-established (e.g. [2]). In contrast speech events are mostly ill-defined. In cases when events can be better specified, they often exhibit a wide variation when the speech signals are collected in different acoustic conditions over a large population of speakers. Another difficulty is that some events are as long as a few seconds, while other can as short as only a few milliseconds. It is therefore a great challenge for us to design high performance speech event detectors.

One way to appreciate the issues related to speech event detection is to compare it with keyword spotting (e.g. [3]). Keywords are usually much longer and more stable to detect than the shorter and more variable speech events that have a tendency to be inserted and deleted. Furthermore, detailed timing in keyword detection is not as critical as event detection because we often need to combine these detected events with highly variable segment boundaries to form higher level evidences. So we can no longer afford having only rough segment information. These two new detection requirements also motivate us to modify existing performance evaluation methods commonly used in keyword spotting to suit speech event detection.

In summary, in order to design effective speech event detectors, three key issues need to be investigated, namely: (1) detector selection; (2) techniques to improve detector performance; and (3) modification of detector evaluation criteria according to properties of the detected events. In this study we design frame and segment based detectors, study their properties in detecting manner of articulation, and propose new mechanism to evaluate detector performance. Our results on the TIMIT database indicate that segment based detectors work better than frame based detectors in several key aspects of speech detector design. We also show that the performance of speech event detection can be significantly boosted with the discriminative training.

2. Detector design

In the following we focus our attention on detecting only manner of articulation attributes [4], namely vowel, fricative, stop, nasal, approximant and silence. The above three issues related to detector design are now addressed in detail.

2.1. Frame and segment based detection

Both frame and segment based event detectors can be used. Frame based detectors can be realized with artificial neural networks (ANNs) [5] as demonstrated in [6]. One advantage with such ANN based detectors is that the output scores can simulate the *a posteriori* probabilities of an attribute given the speech signal. In [7], speech attribute "detectors" for manner and place of articulation were designed using ANNs with multiple outputs. Strictly speaking, these event "detectors" can also categorize each speech frame into one of the competing attributes. A "true" detector should only determine if the current speech frame exhibits the specified attribute or not. We need to group consecutive frames that have detection scores higher than a pre-selected threshold to form detected segments. It is clear that the frame detection scores are likely to fluctuate a great deal, resulting in extra detected segments.

On the other hand segment based detectors can be built by combining frame based detectors, or with segment models, such as hidden Markov models (HMMs) [8], which have already been shown effective for ASR. We train two HMMs, one for a target event, and the other for all other competing events. We used these two models to decode a speech utterance, and the segments that are recognized with the target label mapping to the detected target events. Combined detection and verification has also been proposed in keyword spotting to improve overall performance [9]. In this study, only detection is discussed, verification will be realized in the evidence verifier module in ASAT in the future.

2.2. Performance measurement

Before we describe detector optimization techniques in detail we need to demonstrate the need for new performance evaluation methods. A pseudo example is given below as an illustration. For the fricative event, the reference and detected strings are listed as ref.mlf and det.mlf in the commonly adopted HTK [10] MLF format in the following.

ref.mlf	det.mlf
#!MLF!#	#!MLF!#
"*/si1039.lab"	"*/si1039.rec"
0 8 fricative -3.1	0 13 fricative -13.1
8 13 fricative -10.0	13 21 non -12.1
13 21 non -12.1	21 24 fricative -16.0
21 30 fricative -34.2	24 30 fricative -18.2
30 41 non -19.0	30 41 non -19.0
41 45 fricative -3.0	41 45 fricative -3.0
.	.

In HTK, the detection false alarms (FA) and hits can be obtained with its HResults tool. If the start and end times of a detected event lies in a segment with an identical label in the reference, then the detected segment represents a hit, otherwise an FA is flagged. This evaluation mechanism may reduce the number of actual hits and increase the number of true FAs of shorter speech events when there are contiguous segments with the same label. The HTK tool will report 3 hits and 1 FA for this example, judging the segment (8, 13) in ref.mlf as a false rejection (FR), and the segment (21, 24) in det.mlf as an FA.

Looking into the segments in the two MLF lists more closely the segment (0, 13) in det.mlf is clearly been broken into two segments at (0, 8) and (8, 13) in ref.mlf. In addition, the segment (21, 30) in det.mlf is equivalent to the combined segments at (21, 24) and (24, 30) in ref.mlf. So we need to consider 4 hits and 0 FA for this example. This motivates us to propose new evaluation criteria to more faithfully reflect detector performance. The modified procedure to extend the HTK detection evaluation method is described as follows.

If the reference label lies between the start and end times of a recognized segment with the same label, we regard it as a hit. For every detected segment, we consider it as a FA only if the center of the recognized segment falls into the reference segment with a different label. This modification gives us the correct 4 hits and 0 FA in the above the example.

The FA, FR and error rates are defined as follows:
FA Rate = (# of FAs) / (total # of non-target)
FR Rate = (# of FRs) / (total # of targets)
Error Rate = (# of FAs + # of FRs) / (total # of labels)

Another tool is the DET curve [11] widely used for evaluating speaker verification systems. It is different from the conventional receiver operating characteristic (ROC) [2] curve by using a non-linear scale on both the FA and FR axes to give an easy observation of system contrasts which may be very little and not clearly displayed on the ROC curves. It is not easy to plot the DET curves for the ASAT detection tasks, because the reference and detected target segments are often not aligned. So, we plot the DET curve only when detecting with known segment boundaries. First, we obtain the segments by a forced alignment of the utterance. Second, we retrieve the segment scores from the alignment segments. For the HMM detectors, the segment score is the log likelihood ratio (LLR) score. As for the ANN based detectors, the segment score is: $score = \sum_{t=1}^{T} \log\left[s(O_t)/\left(1 - s(O_t)\right)\right]$, where $s(O_t)$ is the output score for the *t*-th frame. By varying the threshold and comparing it to the segment score, we can decide if the detected segment is the desired target. The FA and FR rates are then computed and used to plot DET curves.

The third utility is the LLR histograms for the target and non-target segments, a conventional way to visualize the FA and FR rates with varying operating points (e.g. [9]). We need a reference label for each detected segment for deciding a hit or miss. For idealized detection with known segment boundaries, we can easily get the segment labels as either a target or non-target event from the reference strings. As for detection with unknown segment boundaries, the situation is more ambiguous. For every detected segment, we assign the label as the one in the reference string that corresponds to the center of the detected segment. For every sample segment, the LLR score is then computed and used to plot the histograms.

2.3. Detector optimization

In order to enhance detector performance, we can use discriminative speech parameters and models with detailed acoustic resolution and refined content dependency. All other algorithms that have been shown to improve ANN and HMM capabilities can also be incorporated. Discriminative training (e.g. [12], [13]) can be used as well. In this study we train detectors with string based MCE (STR_MCE) [12] and segment based MCE (SEG_MCE) [13] criteria to boost performance. Context dependent detectors will be reported in future studies. For MCE training, the LLR is defined as:

$$LLR(O) = \log L(O|\Lambda_0) - \log L(O|\Lambda_1)$$

where $L(O|\Lambda_0)$ and $L(O|\Lambda_1)$ are the likelihoods of target and non-target models. Misclassification measures for the target and non-target models are $d_0(O) = -LLR(O)$ and $d_1(O) = LLR(O)$.

3. Experiment

A phonetically balanced speech corpus is needed for our evaluation. The TIMIT database [14] is chosen and used in all of the following experiments. Excluding utterances used for speaker adaptation (SA), there are a total of 3696 and 1344 utterances in the training and testing sets, respectively.

We design frame based ANN [6] and segment based HMM detectors for the manner of articulation attributes. All ANN detectors share the same structure, with no parameter tuning performed. The input to the networks has 9 frames (the frame rate is 10 msec in the current system) of 12MFCCs + energy, giving a total of 117 input nodes. The hidden layer has 100 nodes. For the current application the output layer has only one node, and its value is 1 if the desired attribute is present at the center frame and 0 otherwise.

The segment based HMM detectors were trained by using HTK. The input features are the same as the above ANN case, and the 117 elements were grouped into a 39-dimensional vector, representing the static MFCC vector, plus their first and second order time derivatives. A pair of target and non-target models was trained for every event of interest. We experimented with both STR_MCE and SEG_MCE detectors. All HMMs are context independent and have 3 states, with each state having 32 Gaussian mixture components.

3.1. Frame versus segment based detection

One problem with the frame based detectors is that it often generates extra noisy segments due to fluctuation of detection scores. Figure 1 compares frame and segment based detectors for the fricative manner. The top panel is a spectrogram. The panel below shows the reference segments. The bottom two panels are detection curves for the frame and segment detectors, respectively. The detected segment, achieved by the segment detector in bottom panel, is more similar to the reference segments. In the third panel for the frame based detector, we get one extra segment due to noisy segments formed if we set a threshold of 0.5 (shown as the dashed line) on the detection curves.

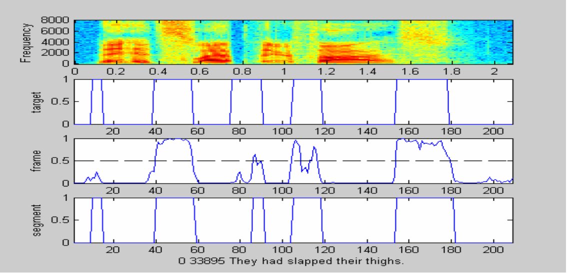

Figure 1 Detection curves of the ANN frame and HMM segment based detectors for the fricative attribute

3.2. Measurement of detector performance

If detection is performed with known segment boundaries, there are no insertion and deletion errors. Using the HTK tool and the proposed measurement discussed in Section 2.2, we get the same result. However, the segment boundaries cannot be known in advance. These two measures will thus result in different error rates. We compare these two measurements using the baseline HMM detectors in the following.

One of the advantages of the new evaluation criterion is to take into consideration the case of merged segments. In the three events listed in Table 1, the fricative manner has the smallest probability of co-occurring in consecutive segments. As a result, the difference between the HTK and new measurement is also the smallest. On the other hand voicing attributes often occur contiguously in an utterance. Our proposed measure gives a much more reasonable detection error rate when compared with the HTK tool. The difference in performance for the vowel manner lies in between.

Error rate (%)	HTK	New
Vowel	5.3	1.7
Fricative	9.0	7.9
Voicing	23.2	2.3

Table 1 Detection errors of the baseline HMM detectors

3.3. Comparison of segment detectors

We measure the performance of segment detectors with and without using information of the segment boundaries.

3.3.1. Detection with known segment boundaries

Detection assuming known segment boundaries is an idealized experimental setup. Given these segments, we can sum over all frame scores of the ANN detector to smooth noisy outputs to some extent. Even after this smoothing, the HMM detectors still outperforms the ANN detector as shown in the DET curve comparison in Figure 2. If it is used in detection with unknown segment boundaries, the noisy nature at the output node of the ANN detectors will make it much worse than the HMM detectors. So in the following experiments, we only compare the performance of the HMM detectors.

We can also see that the SEG_MCE detector outperforms the STR_MCE detector. This is because the SEG_MCE detector only aims at separating the target and non-target models by reducing the substitution errors given the "fixed" boundaries, while the STR_MCE detector attempts to reduce the overall errors, including insertions and deletions of the whole string. As a consequence, the SEG_MCE detector works best with known segment boundaries.

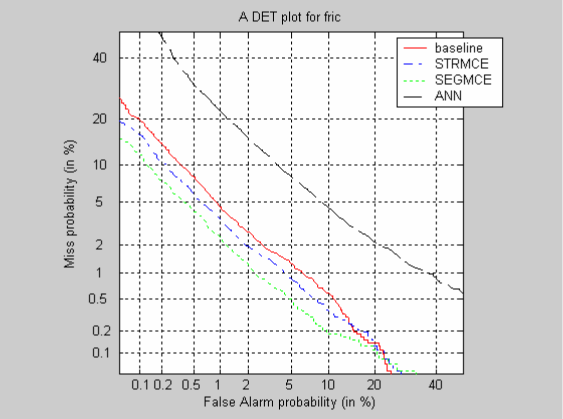

Figure 2 Comparing DET curves for detecting fricative event with baseline, STR_MCE, and SEG_MCE HMM detectors and the ANN detector with known segment boundaries

3.3.2. Detection with unknown segment boundaries

In the actual situation of real-world event detection, segment boundaries cannot be assumed known. We have to use the two competing HMMs to decode each utterance into the target and non-target segments, and compute the FA, FR and error rates accordingly. As shown in Table 2, the MCE optimized detectors work better than the baseline HMM detectors. In the case of vowel and approximant detection, the SEG_MCE detectors were superior to the corresponding STR_MCE detectors. In the other four cases, the SEG_MCE detectors work slightly worse than the STR_MCE detectors. This inconsistence in overall error comparison was caused by not knowing the correct segment boundaries. In summary the SEG_MCE detectors minimizes substitution errors and give better model separation, while the STR_MCE detectors attempt to minimize the overall errors.

Error rate (%)	Baseline	STR_MCE	SEG_MCE
Vowel	1.7	2.4	1.8
Fricative	7.9	4.8	4.9
Stop	9.9	5.3	5.4
Nasal	11.2	5.0	5.4
Approximant	7.3	6.3	5.2
Silence	2.1	0.6	0.8

Table 2 Detection error rates of three segment detectors

It can be seen that the SEG_MCE detector behaved slightly worse than the baseline HMM detector only for the target vowels. However, if both the target and non-target events are considered, the combined error rate of 5.2% for the SEG_MCE detector is still better than the 5.7% error rate obtained with the baseline HMM detector. This can be explained by the MCE training policy that aims at reducing the total number of classification errors of all the competing classes. Therefore there is no guarantee that the target error rate will be reduced as well. So, it is important to develop other discriminative training algorithms that directly reduce the target event error rate and different combinations of FA and FR rates.

Two sets of LLR histograms for the fricative event are plotted in Figure 3 for comparison. The superiority of the SEG_MCE detector over the baseline detector is clearly shown with the target histogram moving to the right, and the non-target (impostor) histogram moving to the left after MCE training. This results in a larger separation and a smaller overlapping region, which also implies a smaller error rate.

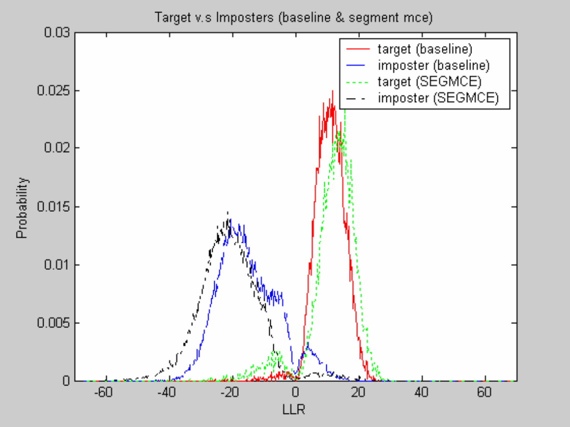

Figure 3 The LLR plot for the baseline and the SEG_MCE detector for fricative event with unknown segment boundaries

4. Conclusion

We have addressed three key issues related to realizing effective speech event detectors. Frame based ANN and segment based HMM detectors were compared. Experimental results on detecting the manner of articulation events using the TIMIT database showed that the HMM detectors often gave much better performance than the ANN detectors. We also found that these detectors can be improved significantly with the discriminative learning. Due to the brief nature of some speech events we also need some improved techniques to accurately evaluate the detector performance.

Some additional research issues are worth pursuing. So far the detectors are all based on 10-msec MFCCs. If optimal speech parameters can be derived for some specific speech events, we expect the corresponding detectors to be optimal. For example, voice onset time has been shown effective in discriminating voices against unvoiced stop sounds. To capture speech events in context we will extend to designing context dependent detectors as well. In addition to the conventional MCE formulation, other criteria to directly minimize different combinations of FA and FR errors need to be investigated. Detectors for higher level event to combine multiple spatial and temporal speech events will also be studied in the ASAT framework in the future.

5. Acknowledgement

Part of this effort was supported under the NSF SGER grant, IIS-03-96848 and NSF ITR grant, IIS-04-27413. The first author also thanks his previous colleagues, Gang Guo, Xiaobing Li and Bo Liu, at the Speech Lab at USTC iFlytek for assisting in detector evaluation tool design.

6. References

[1] Lee, C.-H., “From knowledge-ignorant to knowledge-rich modeling: a new speech research paradigm for next generation automatic speech recognition,” *Proc. ICSLP*, 2004.

[2] Kay, S.M., *Fundamentals of Statistical Signal Processing. Volume II : Detection Theory*, Prentice Hall, 1998.

[3] Rose, R.C., “Keyword detection in conversational speech utterances using hidden Markov model based continuous speech recognition,” *Computer Speech and Language*, vol. 9, no. 4, pp. 309-333, 1995.

[4] Kirchhoff, K., “Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments,” *Proc. ICSLP*, 1998.

[5] Haykin, S., *Neural Networks: a Comprehensive Foundation (2nd edition)*, Prentice Hall, 1998.

[6] Li, J., Tsao, Y. and Lee, C.-H., “A study on knowledge source integration for rescoring in automatic speech recognition,” *Proc. ICASSP*, 2005.

[7] Hacioglu, K., Hacioglu, B., and Ward, W., “Parsing speech into articulatory events,” *Proc. ICASSP*, pp. 925-928, Montreal, Canada, 2004.

[8] Rabiner, L. R., “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. IEEE*, vol. 77, no. 2, pp. 257-286, 1989.

[9] Rahim, M. G., Lee, C.-H., and Juang, B.-H., “Discriminative utterance verification for connected digits recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 266-277, 1997.

[10] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P., *The HTK Book (for HTK Version 3.2)*, Cambridge University, 2002.

[11] Martin, A, Doddington, G., Kamm, T., and Ordowski, M., and Przybocki, M., “The DET curve in assessment of detection task performance,” *Proc. EuroSpeech,* pp. 1895-1898, 1997.

[12] Juang, B.-H., Chou, W., and Lee, C.-H., “Minimum classification error rate methods for speech recognition,” *IEEE Trans on Speech and Audio Processing*, vol. 5, no. 3, pp. 257-265, 1997.

[13] Chou, W., Juang, B.-H., and Lee, C.-H., “Segmental GPD training of HMM based speech recognizer,” *Proc. ICASSP,* pp. 473-476, 1992.

[14] Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscusk, J.G., Pallett, D.S., and Dahlgren, N.L., “DARPA TIMIT acoustic-phonetic continuous speech corpus,” U.S. Dept. of Commerce, NIST, Gaithersburg, MD, February 1993.