A SEGMENTATION POSTERIOR BASED ENDPOINTING ALGORITHM

YanLu XIE[2]*, Yu SHI*[1]*, Frank K. SOONG*[1]*, BeiQian DAI*[2]

[1]Microsoft Research Asia, Beijing, China
[2]MOE-MS Key Laboratory of Multimedia Computing and Communication
University of Science and Technology of China, Hefei, China
xieyl@mail.ustc.edu.cn, {yushi,frankkps}@microsoft.com, bqdai@ustc.edu.cn

ABSTRACT

A segmentation posterior probability based endpointing algorithm for robust ASR is proposed. First, each speech signal is partitioned into homogeneous segments via auto-segmentation. Then posterior probabilities of all possible endpoints are computed, based on the segmentation likelihoods of all levels in a selected range. Endpoints with the highest posterior probabilities are finally selected. The new method differs from the previous auto-segmentation and clustering based algorithm on that the former considers hypotheses from several levels, while the latter depends only on one appropriate level. Another potential benefit of the proposed method is that any endpointing or VAD results can be integrated, as hypotheses, into the posterior probability framework. Experiments based on the AURORA2 digit database show the robustness of the proposed method.

Index Terms— Speech recognition, endpointing, VAD, auto-segmentation, posterior probability

1. INTRODUCTION

Endpointing or voice activity detection (VAD) is a key component in speech recognition systems. First, in order to improve the performance of the speech recognition systems under adverse environments, various noise reduction algorithms have been proposed. In fact, most of these algorithms especially the single microphone based methods such as Wiener filtering (WF) and Spectral Substation (SS) often require an estimate of the noise statistics by means of a precise VAD. Second, frame dropping (FD) is frequently used to reduce the number of insertion errors in speech recognition. Since it is based on the VAD, speech frames incorrectly labeled as silence causes unrecoverable deletion errors, and silence frames incorrectly labeled as speech could increase the insertion errors.

Unfortunately the performance of most of the current VAD algorithms decreases greatly when the background noise is high. Thus a lot of researchers focus on the lower SNR environment. Methods being proposed include higher-order statistics [1], combinations of different features [2], using long-term information [3], using order statistic filters [4], and so on. The European Telecommunication Standards Institute (ETSI) also approved the new standard for feature extraction and distributed speech recognition (DSR) in 2002. The advanced front-end (AFE) [5] block in the standard proposed different VAD methods for Wiener filtering speech enhancement (WF AFE VAD) and non-speech frame dropping (FD AFE VAD).

An auto-segmentation based endpointing algorithm is proposed by Yu *et al* recently [6][7]. The algorithm consists of two successive steps: (1) homogeneous segment partitioning and (2) segment clustering. In the first step, the algorithm divides a time series into homogeneous partitions via a level building dynamic programming (DP). Then the optimal level is found according to the segmentation homogeneity penalized by segmentation complexity. The second step groups the segments into two clusters: speech and background noise, and finally endpoints are easily obtained. The algorithm considers long-term information and outperforms the AFE VADs under lower SNRs. However the method only chooses the optimal level and neglects other levels' information which is also helpful to determine the endpoints in some sense.

In this paper a posterior probability method is proposed to incorporate the useful information from other levels. With the segmentation likelihood of each level, endpoints can be found with the maximum posterior probability criterion. And more important, the method has a potential to integrate any endpointing or VAD results into the posterior probability framework. Experiments based on the AURORA2 database show that the method is better than the AFE VADs and is comparable to the old clustering based method in terms of speech recognition performance.

2. AUTO-SEGMENTATION AND LEVEL BUILDING

For a given time interval $I = \{n, n = 1, \ldots, N\}$ which contains N frames of speech and a predefined parameter L ($1 \leq L \leq N$) which represents the total number of segments to be produced, segmentation $\mathcal{S}(N, L)$ is defined as a set of L blocks

$$\mathcal{S}(N, L) = \{S_k, k = 1, \ldots, L\} \tag{1}$$

where each block is a set of frames represented by consecutive time indices $S_k = \{n_{k-1}+1, \ldots, n_k\}$ satisfying the nonoverlapping and nonskipping conditions $\bigcup_k S_k = I$ and $S_k \bigcap S_{k'} = \emptyset$ if $k \neq k'$. Here n_k is the end frame of segment S_k.

The segmentation homogeneity criterion is defined as

$$H(N, L) = \sum_{k=1}^{L} D(n_{k-1} + 1, n_k) \tag{2}$$

where $D(n_{k-1} + 1, n_k)$ indicates a measure function of homogeneity associated with segment k positioned from frame $n_{k-1}+1$ to n_k, which is defined as the within-segment distortion as in [6] and [7]. An optimal segmentation $\mathcal{S}^*(N, L)$ can be obtained by minimizing $H(N, L)$ over all segment boundaries:

$$\mathcal{S}^*(N, L) = \arg\min_{\mathcal{S}, |\mathcal{S}|=L} H(N, L) \tag{3}$$

$$H^*(N, L) = \min_{\mathcal{S}, |\mathcal{S}|=L} H(N, L) \tag{4}$$

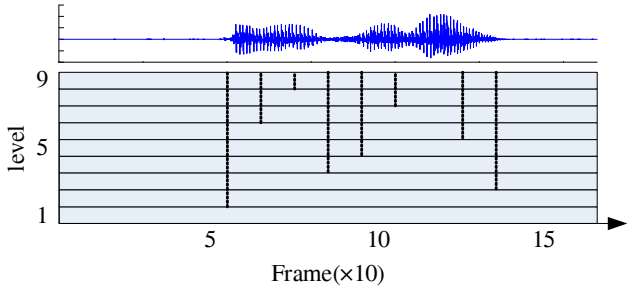
The search procedure is implemented similar to the level building DP algorithm [8], i.e., the lth level has l segments. For a particular frame $n, 1 \le n \le N$, and level $l, 1 \le l \le L$, defining $H^*(n,l)$ as the distortion of the optimal partition up to frame n at level l and $m^*(n,l)$ as the ending location of the next-to-last segment of the optimal partition up to frame n at level l, we have the recurrence equation as:

$$H^*(n,l) = \min_{l-1 \le j < n} \{H^*(j, l-1) + D(j+1, n)\} \quad (5)$$

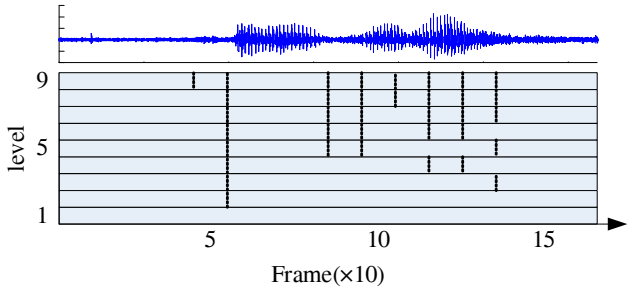
$$m^*(n,l) = \arg\min_{l-1 \le j < n} \{H^*(j, l-1) + D(j+1, n)\} \quad (6)$$

The optimal segment boundaries at level l will be obtained via tracing back from $n_l = N$ using back pointers led by $m^*(N,l)$.

The level building process will repeat until the maximum level L is reached. Fig. 1 shows two examples of the result. The vertical dashed lines denote segment boundaries at each level. From the figure, we will see that the endpoints at certain levels are very stable and consistent, even if the speech is corrupted by noise.



(a) clean



(b) subway noise, SNR=10dB

Fig. 1. Level building examples.

3. POSTERIOR PROBABILITY BASED ENDPOINTING

For a given level l and segment $k, 1 \le k \le l$, we can obtain the centroid $\vec{\mu}_{l,k}$ and covariance matrix $\Sigma_{l,k}$ from the level building results. Thus the likelihood for the nth frame, $n_{l,k-1}+1 \le n \le n_{l,k}$, of d-dimensional feature vector $\vec{\mathbf{x}}_n$ belonging to the segment is calculated as:

$$P(\vec{\mathbf{x}}_n | l,k) = \mathcal{N}(\vec{\mathbf{x}}_n; \vec{\mu}_{l,k}, \Sigma_{l,k}), n_{l,k-1}+1 \le n \le n_{l,k} \quad (7)$$

where $\mathcal{N}(\vec{\mathbf{x}}; \vec{\mu}, \Sigma)$ denotes a normal distribution with mean $\vec{\mu}$ and covariance matrix Σ, $n_{l,k}$ denotes the end frame of the segment. In this paper, we assume that the covariance matrix is an identity matrix, i.e., $\Sigma_{l,k} = \mathrm{I}$. Then we have the likelihood of the segment (l,k) and the likelihood of the whole speech signal in level l as:

$$P(\mathbf{X}_{n_{l,k-1}+1}^{n_{l,k}} | l,k) = \prod_{n=n_{l,k-1}+1}^{n_{l,k}} P(\vec{\mathbf{x}}_n | l,k) \quad (8)$$

$$P(\mathbf{X}_1^N | l) = \prod_{k=1}^{l} P(\mathbf{X}_{n_{l,k-1}+1}^{n_{l,k}} | l,k) \quad (9)$$

From Fig. 1 we found that the endpoints of different levels are quite similar to N-best list, which is often used in post-processing of ASR systems and contains the N most likely hypotheses generated by a preliminary pass of search. In ASR systems, given the N-best output, the posterior probability of a specific word can be estimated by summing up the posterior probabilities of all string hypotheses that contain the word with the same starting and ending time [9]. Similarly, in the proposed endpointing algorithm, for a certain endpoint, its posterior probability can be estimated by summing up the posterior probabilities of a selected range of levels ($l_{\min} \le l \le l_{\max}$) that contain the similar endpoint:

$$P(n_{\text{start point}} | \mathbf{X}_1^N) = \frac{\sum_{l=l_{\min}}^{l_{\max}} P(\mathbf{X}_1^N | l)^\alpha}{\sum_{l=l_{\min}}^{l_{\max}} P(\mathbf{X}_1^N | l)^\alpha} \quad \text{if } n_{l,1} \approx n_{\text{start point}} \quad (10)$$

$$P(n_{\text{end point}} | \mathbf{X}_1^N) = \frac{\sum_{l=l_{\min}}^{l_{\max}} P(\mathbf{X}_1^N | l)^\alpha}{\sum_{l=l_{\min}}^{l_{\max}} P(\mathbf{X}_1^N | l)^\alpha} \quad \text{if } n_{l,l-1}+1 \approx n_{\text{end point}} \quad (11)$$

where Equation (10) is for start point and Equation (11) is for end point. α denotes an exponential weight and can be trained on a training set.

In the above equations, $l_{\min}$ and $l_{\max}$ are parameters having to be decided. Experiments show that $l_{\min} = 4$ can get satisfied results. However $l_{\max}$ affects the posterior probability much more. First, $l_{\max}$ should be large enough to take levels with correct endpoint into the hypothesis list. On the other hand, $l_{\max}$ should not be too large, otherwise many incorrect hypotheses which have higher posterior probabilities may be taken into consideration. This will do harm to the endpoint detection.

Intuitively we think that $l_{\max}$ is related to both speech length and segmentation complexity. Therefore a penalty related to them is added to the optimal segmentation score at each level to balance the homogeneity and complexity:

$$F(N,l) = H^*(N,l) + \lambda P(N,l), 1 \le l \le L \quad (12)$$

where λ is a penalty weight and

$$P(N,l) = \#(N,l) \log(N) \quad (13)$$

denotes the penalty term, and $\#(N,l)$ indicates the number of parameters in the segmentation. For minimum segmentation distortion criterion and d-dimensional feature vector, it can be calculated as $\#(N,l) = l \times d$. Then $l_{\max}$ can be obtained by minimizing $F(N,l)$ over all levels:

$$l_{\max} = \arg\min_{1 \le l \le L} F(N, l) \tag{14}$$

The penalty weight is chosen to minimize the mean square error of the estimated $l_{\max}$ over the training set as:

$$\lambda^* = \arg\min_{\lambda} \sum_{i \in \text{training set}} (l_{\max}(i) - l_i)^2 \tag{15}$$

where l_i denotes the maximum level whose endpoint is correct associated with the ith sample.

4. EXPERIMENTAL RESULTS

Receiver operating characteristic (ROC) curves and speech recognition results based on the AURORA2 database were used to verify the performance of the proposed algorithm. ($\alpha = 0.006$, $\lambda = 0.32$) and ($\alpha = 0.0002$, $\lambda = 0.1105$) are parameters for start and end point detection, respectively.

4.1. Receiver operating characteristic curves

First, the endpoint detection method was evaluated by means of the ROC curves, which can completely describe endpointing performance. The non-speech hit-rate (HR_0) and speech hit-rate (HR_1) are defined as the fraction of all actual pause or speech frames that are correctly detected as pause or speech frames, respectively:

$$\mathrm{HR}_0 = \frac{\mathrm{C}(0|0)}{\mathrm{C}_{\mathrm{ref}}(0)}, \quad \mathrm{HR}_1 = \frac{\mathrm{C}(1|1)}{\mathrm{C}_{\mathrm{ref}}(1)} \tag{16}$$

where $C_{\mathrm{ref}}(0)$ and $C_{\mathrm{ref}}(1)$ are the counts of real non-speech and speech frames in the whole database, respectively, while $C(0|0)$ and $C(1|1)$ are the counts of non-speech and speech frames correctly classified. The "real" speech frames and "real" speech pauses were determined by aligning clean test data to a set of HMM models trained on clean data in both training and test sets in the database.

HR_0 and FAR_0 ($\mathrm{FAR}_0 = 100 - \mathrm{HR}_1$), the non-speech false-alarm rate, were determined in each noise condition for the proposed endpointing algorithm. HR_0 as a function of FAR_0 for different hangovers is shown in Fig. 2. (Hangover is the appended time duration to the period in which voice activity is detected. It is commonly used in voice activity detector to produce an extended voice detection period to avoid extending noise spikes.) The results are averaged values over all noise types and SNR levels. Operating points of the AFE VADs and operating point chosen for speech recognition are also included. In the figure, each solid line represents one start point hangover, while each dotted line represents one end point hangover. Different colors represent different hangover values, i.e., color blue, green, red, cyan and magenta represent hangovers of 0.2, 0.15, 0.1 0.05, 0.0 sec, respectively. It can be derived from the figure that the posterior probability based endpointing yields a compromise between the low FAR0 and the low HR0 as to FD AFE VAD and WF AFE VAD. Since frame dropping affect the speech recognition much, the low FAR_0 is more important. Thus the operating point chosen for speech recognition usually has a slight lower HR_0 a much lower FAR_0 than WF AFE VAD.

4.2. Influence of endpoint detection on speech recognition

As illustrated before, noise detection is playing two important roles in speech recognition in adverse environments. In noise reduction, since noise parameters such as its spectrum are updated during non-speech periods, a good noise detection algorithm is critical for an effective estimation of noise that is required by speech enhancement systems. On the other hand, non-speech frame dropping is strongly influenced by the performance of the detector in effectively reducing the number of insertion errors caused by the noise but not leading to too many irrecoverable deletion errors caused by speech misclassification errors. Thus, an effective endpointing algorithm for robust speech recognition needs a compromise between speech and non-speech detection accuracy.

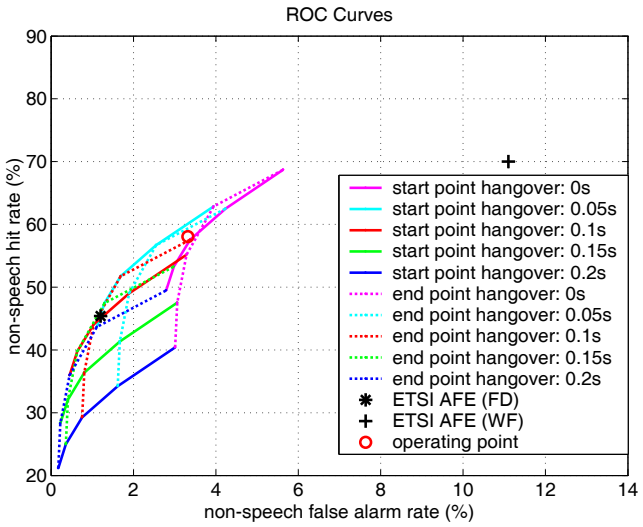


Fig. 2. *ROC curves.*

The recognition experiments were performed based on the AURORA2 database [10]. Since few silences exists between speeches, endpoint detection is enough for noise estimation and noise frame dropping. The reference front-end (baseline) is what was used in the ETSI AURORA project for DSR [11]. The AFE features were extracted by means of the ETSI software [12]. We only used clean training in our analysis.

In order to compare the proposed method to the AFE standard, the VADs of the full standard (including both the noise estimation VAD and frame dropping VAD) were replaced by the proposed endpointing method. Results associated with the HMM based endpoints (reference) and auto-segmentation and clustering based endpoints [7] were also provided. All results were averaged over the three test sets of the AURORA2 recognition experiments. More clearly, the experiment structure is:

1. replace the WF VAD of the AFE standard and do not perform frame dropping
2. replace the FD VAD of the full AFE standard
3. replace both WF and FD VADs of the full AFE standard

The same feature extraction scheme was used for both training and testing. If FD is utilized, exact speech periods expanded by a small time duration, rather than only the exact speech periods as in [7], were kept and consequently, all the frames out of the periods were discarded. The expanded minor duration is necessary for silence model training.

Table 1 to 3 exhibit all recognition results in the clean training. In the tables, "AC" means the auto segmentation and clustering based endpointing method, while "PP" represents the proposed

posterior probability based method. Note that AFE standard uses different VADs for noise suppression and frame dropping. Table 1 demonstrates the performance of noise reduction scheme in robust speech recognition. The four methods have a similar performance. PP performed even a little bit better than REF. This may due to the inaccuracy of the HMM based endpoints. Table 2 shows the recognition results of the full AFE standard and the modified standard via only replacing the FD VAD by others. The word error rate was reduced from 14.0% to 13.5% when AC was used and 12.8% word error rate was got when PP was used. This shows the effectiveness of the posterior probability based algorithm. Table 3 shows the experimental results of the full AFE standard and the modified standard via replacing both WF and FD VADs by others. Both REF and AC achieved an absolute improvement of 0.4%, while PP got the similar results as in Table 2. However, the relative improvement of PP is still about 8.6% as to AFE, and PP still performed better than AC.

Table 1. *Comparison of four methods for noise suppression in AFE.*

System	AFE without FD			
VAD (WF)	REF	AFE	AC	PP
Clean	99.1	99.1	99.1	99.1
20 dB	98.1	98.0	98.0	98.0
15 dB	96.6	96.4	96.5	96.5
10 dB	92.6	92.3	92.5	92.8
5 dB	82.4	82.2	82.2	82.7
0 dB	58.2	58.0	57.9	59.1
-5 dB	27.4	26.9	27.2	28.3
Avg. (0-20 dB)	85.6	85.4	85.4	85.8

Table 2. *Comparison of four methods for frame dropping in AFE.*

System	full AFE			
VADs (WF/FD)	AFE/REF	AFE/AFE	AFE/AC	AFE/PP
Clean	99.2	99.2	99.2	99.3
20 dB	98.3	98.1	98.3	98.4
15 dB	97.0	96.7	96.9	97.1
10 dB	93.9	92.8	93.3	93.9
5 dB	85.1	82.9	83.4	84.6
0 dB	63.3	59.6	60.8	61.8
-5 dB	32.2	27.6	29.0	29.4
Avg. (0-20 dB)	87.5	86.0	86.5	87.2

Table 3. *Comparison of four methods for full AFE.*

System	full AFE			
VADs (WF/FD)	REF/REF	AFE/AFE	AC/AC	PP/PP
Clean	99.3	99.2	99.3	99.4
20 dB	98.4	98.1	98.3	98.4
15 dB	97.1	96.7	97.0	97.2
10 dB	94.1	92.8	93.5	93.8
5 dB	85.3	82.9	83.8	84.6
0 dB	64.4	59.6	61.7	61.8
-5 dB	33.5	27.6	30.5	29.8
Avg. (0-20 dB)	87.9	86.0	86.9	87.2

5. CONCLUSION AND FUTURE WORK

In this paper, we proposed a robust endpoint detection algorithm based on auto segmentation and maximum posterior probability criterion. Due to the self-segmentation nature, the approach does not need any noise models, and long-term information extracted from the segments can be used. On the other hand, since posterior probability was used, multiple endpoint hypotheses produced by the level building dynamic programming, or potentially by other endpointing and VAD algorithms, can be involved. Thus, the method is more robust to noise. Experiments on the Aurora2 digit database showed that the proposed method outperformed the AFE standard VADs when used for WF, FD, and both of them. The reduction of the word error rate was 8.6% over AFE VADs and 2.3% over previous auto-segmentation and clustering based method.

During the experiments, we noticed that l_{max} is a key parameter in the framework. How to estimate it accurately is not trivial. The future work would be discovering an efficient method to select it.

6. REFERENCES

[1] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the lpc residual domain," *IEEE Trans. SAP.*, vol. 9, no. 3, pp. 217–231, May 2001.

[2] S. G. Tanyer and H. Ozer, "Voice activity detection in nonstationary noise," *IEEE Trans. SAP.*, vol. 8, no. 4, pp. 478–482, July 2000.

[3] J. Ramírez, J. C. Segura, C. Benítez, Á. d. l. Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, pp. 271–287, 2004.

[4] J. Ramírez, J. C. Segura, C. Benítez, Á. d. l. Torre, and A. Rubio, "An effective subband OSF-based VAD with noise reduction for robust speech recognition," *IEEE Trans. SAP.*, vol. 13, no. 6, pp. 1119–1129, November 2005.

[5] ETSI ES 202 050, *Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms*.

[6] Y. Shi, F. K. Soong, and J.-L. Zhou, "Auto-segmentation based partitioning and clustering approach to robust endpointing," in *ICASSP2006*, May 2006, vol. I, pp. 793–796.

[7] Y. Shi, F. K. Soong, and J.-L. Zhou, "Auto-segmentation based VAD for robust ASR," in *Interspeech2006-ICSLP*, September 2006, pp. 1958–1961.

[8] C. S. Myers and L. R. Rabiner, "A level building dynamic time warping algorithm for connected word recognition," *IEEE Trans. ASSP.*, vol. 29, pp. 284–297, 2 1981.

[9] R. Schwartz and Y. L. Chow, "The N-best algorithm: an efficient and exact procedure for finding the N most likely sentence hypotheses," in *ICASSP1990*, 1990, pp. 81–84.

[10] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA ITRW ASR*, 2000, pp. 181–188.

[11] ETSI ES 201 108, *Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms*.

[12] ETSI ES 202 212, *Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Extended advanced front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm*.