ROBUST PARAMETERS FOR AUTOMATIC SEGMENTATION OF SPEECH

A. K. V. SaiJayram V. Rarnasubramanian T. V. Sreenivas

Department of Electrical Communication Engineering
Indian Institute of Science, Bangalore 560 012, India
email: tvsree@ece.iisc.ernet.in

ABSTRACT

Automatic segmentation of speech is on important problem that is useful in speed recognition, synthesis and coding. We explore in this paper: the robust parameter set, weightingfunction and distance measure for reliable segmentation of noisy speech. It is found that the MFCC parameters, successful in speech recognition. holds the best promise far robust segmentation also. We also explored a variety of symmetric and asymmetric weighting lifters, from which it is found that a symmetric lifter of the form $1 + A\sin^{1/2}(\pi n/L)$, $0 \leq n \leq L-1$, for MFCC dimension L, is most effective. With regard to distance measure. the direct L_2 norm is found adequate.

1. INTRODUCTION

Segmentation techniques, by and large, aim at producing phonetically and sub-phonetically segmented speech in an acoustically consistent manner. When an acoustic definition of phoneme-like units is required, manual segmentation and labeling becomes tedious, time consuming and, acoustically inconsistent. Different automatic segmentation techniques have been proposed to overcome the problem associated with manual segmentation, Automatic segmentation techniques broadly fall under the following categories [4]: (i) spectral transition measures, ii) maximum likelihood segmentation, iii) temporal decomposition, iv) scale-space filtering, v) multi-level segmentation, vi) variable length segment quantization procedure and, vii) segmental K-means procedure.

Segmentation techniques, regardless of being manual or automatic, play a crucial role in numerous applications which require a large corpus of phonetically segmented speech; one of this is the generation of large acoustically consistent segment corpus as the first stage in the training and design of acoustic subword unit (ASWU) based speech recognition systems [1].

Our focus, in this paper, is on two basic segmentation techniques suitable for ASWU systems, namely, i) spectral transition measure (STM) based segmentation and, ii) maximum likelihood (ML) segmentation [3]. Of these, while STM offers simplicity, the ML segmentation has gained sufficient acceptance in practically most ASWU based speech recognition systems reported so far [1]. Despite this, there has been no consolidated study of these segmentation techniques with respect to the choice of parametric representations, and their relative robustness to additive noise.

The main aim ofthe present paper is to study the performance robusmess ofthe STM and ML segmentation, far different parametric representations. In the study ofthe relative robustness of the parameters, we mainly wish to identify the parametric representations that are more robust to additive noise under different SNR's.

There are several linear-prediction (LP) based parametric representations which are non-linearly related and hence could give rise to a different performance. Also, in any parametric representation, the chosen distance measure and weighting functions play a crucial role in their performance. In addition to these, since MFCC parameters have proved successful for speech recognition, we have studied MFCC with different types of lifters and distance measures far automatic segmentation,

2. SPECTRAL TRANSITION MEASURE

Let $\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_p(t)]'$ be a parameter vector representing spectral features of the signal at lime t. The spectral gradient or a spectral transition measure is given by the derivative of $\mathbf{x}(t)$, $\mathbf{x}'(t)$ as $\mathbf{x}'(t) = \frac{\partial \mathbf{x}(t)}{\partial t}$. Though defined for a continuous case, the derivative $\mathbf{x}'(t)$ can be viewed as a vector-valued distortion between infinitesimally small contiguous acoustic vector sequence. The magnitude of $\mathbf{x}'(t)$, $||\mathbf{x}'(t)||$, is a scalar measure and represents the rate at which the spectral feature vector $\mathbf{x}(t)$ is changing at time t. This scalar measure (derivative vector magnitude) is expected to be large at the boundaries between successive quasi-stationary speech sounds, corresponding to a transition; these are at the instances of sharp rate of change of the vector trajectory in the p dimensional parameter space representing rapid vocal tract changes.

The acoustic observation is a discrete-time vector sequence, given by $\mathbf{x}(n) = [x_1(n), x_2(n), \ldots, x_p(n)]'$, $n = 1, \ldots,$ T, representing a parameter vector $\mathbf{x}(n)$ at discrete time instance or frame 'n'. Here, the derivative $\mathbf{x}'(n)$ is approximated by differences, The successive difference estimate of $\mathbf{x}'(n)$ could be very noisy because of errors due to the parameter estimation process. Hence a low order polynomial fit is used to estimate the trajectory of each vector element. The gradient is obtained by mmse fitting of a straight line to each of the vector elements, within a defined time window. This leads to the slope estimate given by [3],

$$\Delta x_m(n) = \frac{\sum_{k=-K}^{K} k h_k x_m(n+k)}{\sum_{k=-K}^{K} h_k} \tag{1}$$

where, h_k is a symmetric window of length (2K + 1). A sufficiently good estimate of the spectral transition can be obtained by using the above gradient estimates. A scalar measure far the spectral variation (or the spectral derivative magnitude $||\mathbf{x}'(t)||$) is defined as,

$$d_{\Delta}(n) = \sum_{m=1}^{p} (\Delta x_m(n))^2 \tag{2}$$

This is a running estimate of gradient vector norm and it tends to exhibit peaks at the boundaries between speech sounds corresponding to changing vocal tract configuration and is hence a

0-7803-7402-9/02/$17.00 82002 IEEE

measure of its non-stationarity. The problem of finding segment **boundaries,** between two **stationary** segments, **thus** reduces to **a** peak-picking procedure **on** $d_\Delta(n)$, n $= 1,\ldots,$T; the segment boundaries are set to be at the instances of maximum spectral change, i.e. maximum $d_\Delta(n)$.

3. MAXIMUM LIKELIHOOD SEGMENTATION

A speech utterance is given by $X_1^T = (\mathbf{x}_1, \mathbf{x}_2, \ldots, x_T)$, which is **a** discrete observation sequence of T speech **frames,** where, x_n is a p-dimensional acoustic vector at frame 'n', denoted by $\mathbf{x}_n = [x_1(n), x_2(n), \ldots, x_p(n)]'$. A partial sequence extending from frame i to **frame** j is denoted by $X_i^j = (x_i, x_{i+1}, \ldots, x_j)$. The segmentation problem is to find 'm' consecutive segments in the observation sequence X_1^T. Let the segment boundaries be denoted by the set of integers $B = \{b_o, b_1, \ldots, b_m\}$. The i^{th} segment starts at frame $b_{i-1} + 1$ and ends at frame b_i; the beginning and end paints of the sampled speech data are given and fixed, i.e., $b_0 = 0$, and $b_m =$ T.

The main criteria to be satisfied in the segmentation problem is to obtain segments which exhibit maximum acoustic homogeneity within their boundaries. The amount of acoustic inhomogeneity of **a** segment can be measured in terms of an '**intra-segmental** distortion'. It can be measured **as** a sum ofdistances from the frames that span the segment, to the centroid of **these** frames comprising **the** segment. Alternatively, all the frames included in **a** segment can be assumed to have been generated by the same auto-regressive (AR) model, and the intra-segmental distortion can be computed **as** an overall likelihood ratio distortion of these **frames.** The general approach then, is to obtain **a** segmentation with **the** minimum sum of intra-segment distortion. The segmentation is thus **a** problem offinding segment boundaries $\{b_0, b_1, \ldots, b_m\}$ that minimize the **total distortion**

$$D(m,T) = \sum_{i=1}^{m} \sum_{n=b_{i-1}+1}^{b_i} d(\mathbf{x}_n, \mu_i) \qquad (3)$$

where, $D(m,T)$ is the total distortion of a m · segment segmentation of $X_1^T = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$; μ_i is the generalized centroid of the i^{th} segment consisting ofthe spectral sequence $\mathbf{X}_{b_{i-1}+1}^{b_i} = \{\mathbf{x}_{b_{i-1}+1}, \ldots, \mathbf{x}_{b_i}\}$ for a specific distance measure $d(\cdot,\cdot)$. The centroid **can** be viewed **as a** maximum-likelihood estimate of the frames in the segment $\mathbf{X}_{b_{i-1}+1}^{b_i} = \{\mathbf{x}_{b_{i-1}+1}, \ldots, \mathbf{x}_{b_i})$ (under **the** assumption that the **frames** in the segment **are** modeled by a multivariate Gaussian, whose mean μ_i is the centroid being estimated) **as,**

$$\mu_i = \arg\min_{y} \left[\frac{1}{b_i - b_{i-1}} \sum_{n=b_{i-1}+1}^{b_i} d(\mathbf{x}_n, \mathbf{y}) \right] \qquad (4)$$

For the Euclidean distance 'd', μ_i is the average of the frames in the segment $\mathbf{X}_{b_{i-1}+1}^{b_i}$.

The segment boundaries can be solved efficiently using a dynamic programming (DP) procedure. The i^{th} intra-segment distortion is given by

$$\Delta(b_{i-1} + 1, b_i) = \sum_{n=b_{i-1}+1}^{b_i} ||\mathbf{x}_n - \mu_i|| \qquad (5)$$

Let the minimum accumulated distortion upto **the** i^{th} segment (which ends in frame b_i) be denoted **as** $D(i, \mathrm{h})$, i.e., $D(i, b_i)$ is the minimum distonion of **a** segmentation of $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{b_i})$ into i segments. The dynamic programming problem is to find the minimum of

$$D(i, b_i) = \min_{b_{i-1}} [D(i-1, b_{i-1}) + \Delta(b_{i-1} + 1, b_i)] \qquad (6)$$

for all possible b_{i-1}. The segmentation problem is then one of obtaining the minimum **total** distortion $\min\{D(m,$ T$)\}$ (3). **This** is computed efficiently by a trellis realization and the optimal segmentation boundaries $(b_0, b_1, \ldots,$ b,) **are** found by backtracking **on** the trellis after the optimal alignment path is determined corresponding to $\min\{D(m,$ T)).

Likelihood-ratio distortion

A different intra-segment distortion results if each segment is treated as generated by the ***Same*** auto-regressive (AR) model [3]. Here, the DP recursion (6) **takes** the form

$$D(i, b_i) = \min_{b_{i-1}} [D(i - 1, b_{i-1}) + \Delta_{\mathrm{LR}}(b_{i-1} + 1, b_i)] \qquad (7)$$

$$\Delta_{\mathrm{LR}}(b_{i-1} + 1, b_i) = \sum_{n=b_{i-1}+1} \frac{\mathbf{a}_i' R_n \mathbf{a}_i}{\sigma_n^2} \qquad (8)$$

The right hand side of (8) is the likelihood-ratio (LR) distortion between the speech frames $\{\mathbf{x}_n, \mathrm{n} = b_{i-1} + 1, \ldots, b_i\}$ in segment i and **the** corresponding LPC vector $\mathbf{a}_i$, representing the AR model from which frames in segment i are modeled to be generated. Here, $\mathbf{a} = [1, \mathrm{al}, a_2, \ldots, a.]'$ ***is the*** linear prediction ***in***verse filter (assuming speech to be stationary, p^{th} order AR process with **a** white Gaussian, $N(0, \sigma_n^2)$ innovation process) and R_n is the covariance matrix of $\mathbf{x}_n$. The prediction coefficients $\{\mathbf{a}_i\}$ are obtained by first averaging the auto-correlation coefficients of $\{\mathbf{x}_n, \mathrm{n} = b_{i-1} + 1, \ldots,$bi} and obtaining the prediction coefficients corresponding to the average of the autocorrelation coefficients. The boundaries b_i are found by minimizing (7), for **the** likelihood ratio (LR) distortion (8), for all possible segmentations of the observation sequences using (8).

4. EXPERIMENTS AND RESULTS

Database and analysis parameters

We have performed all segmentation experiments on the TIMIT database; here, the speech data is sampled at **16** kHz sampling rate and has an **SNR** of 36dB, **on an** average with respect to background noise in the silence regions The parameters used for thresholding in STM based segmentation and termination conditions in ML **segmentation** are obtained from ***50*** sentences (referred here **as** training set) **taken** from 10 speakers. These parameters **are** then used on test data for evaluating the performance of segmentation using STM and ML; **the test set** is also of **50** sentences, **taken** from 10 speakers. The coefficient order for LPC based features and MFCC are fixed to be 16, with an analysis frame of **20** msec and frame *shift* of 20 msec. Each frame of speech is first pre-emphasized by $1 - 0.95z^{-1}$ and **then** windowed by a Hamming window. The pre-emphasized and windowed frame is then used for parameter estimation.

Feature representation

The different LP related parametric representations used as features, are linear prediction coefficients(LPC), LP cepstral coefficients ***(CEPS)***,log-area ratios (LAR), reflection coefficients(RC) and line-spectral pain (LSP). **We** have omitted **the** last one for the

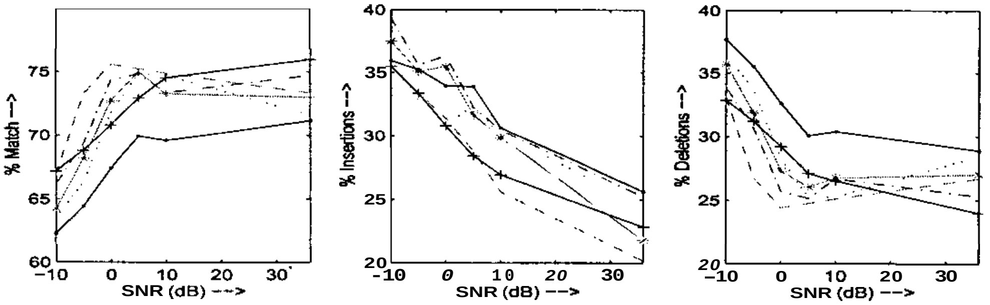

Fig. **1. STM** segmentation: (%M, **%I,** %D) for differentparametric representations for **clean** speech *(36* **dB)** and noisy speech **with** SNRs -10 dB, **-5** dB, 0 dB, **5** dB, 10 dB. Legend •:LPC, · ..:CEPS, ı: LAR, −·:RC, −−: MFCC, +:MFCC-L

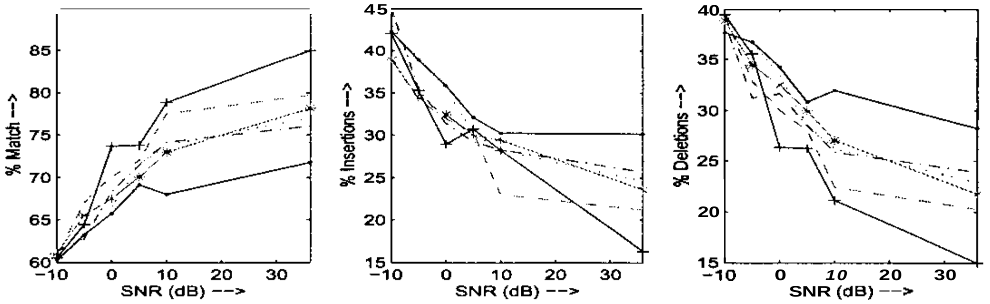

Fig. **2. ML** segmentation: (%M, %I, %D) for differentparametric representations for clean speech *(36* dB) and noisy speech with SNRs -10 dB, **-5** dB, 0 dB, **5** dB, 10 dB. Legend **•:LPC,** · ..:CEPS, ı: LAR, −·:RC, −−: MFCC, +:MFCC-L

present **because** it requires **a** perceptually motivated distance measure for effective performance. Instead, we have included mel-Frequency cepstral coefficients (MFCC) and MFCC with **a** variety of lifters (MFCC-L), because of their effectiveness in noisy speech recognition.

Far MFCC with lifter (MFCC-L), we have explored a class of symmetric windows given by $1 + A\sin^r(\pi n/L)$, $0 \leq n \leq L-1$, for MFCC dimension L; the amplitude A and, the power of $\sin(\pi n/L)$, r, are treated **as** variables for maximizing the segmentation match (%M). The peak amplitude is varied from 1 to L for MFCC of dimension L = 16 and the shape of the lifter is changed by varying r **from 0.25 to 4.** Another **class** of asymmetric triangular Bartlett windows **were** explored by moving the peak position from left to right, keeping the window width fixed. Of all these set of lifters, we found the symmetric lifter to perform best. The optimal lifter is found to be for A $= 4$ and $r = 0.5$. It is found that this optimum lifter performs better than the regular raised-sine lifter [2], providing **4-5%** improvement in the segmentation accuracy.

Distance measures

For ML segmentation with intra-segment distortion, **as** in (5)), we have considered six distance measures, viz., L_1, L_2 (unweighted Euclidean distance (5)), L_3 norms. weighted Euclidean distance (WED: L_2 norm using inverse variance), Mahalanobis distance (MD) and likelihood ratio (LR) (8) far the LPC feature set.

Parameters in STM and ML

Far **the** spectral transition measure **(STM)** based segmentation, the threshold for peak-picking was set as follows: **a** set of ***50* sentences** is used as a training data; for each of these sentences, the threshold was varied within the range of $d_\Delta(n)$ **(2)** and the optimal threshold is obtained as that value for which the peak-picking procedure gives the number of segments same as that of the actual number of segments in **the** manual segmentation of TIMIT. **A** single threshold value is obtained as the average of **the** individual thresholds **over** the 50 training sentences. The window length parameter 'K' in (1) is optimized for maximum segmentation match with manually segmented data and the optimal value of $K = 1$ (window length of $2K + 1 = 3$) **was** determined and set far **all** subsequent segmentations on test data.

In the ML segmentation, **the** segmentation performed using the DP **recursion** (6), (7) can segment the input speech into **a** pre-specified number of segments **or** it can be terminated when the reduction in **the** total distortion *(3)* falls below **a** pre-set threshold. This is the same as the 'average likelihood (or distortion) per frame' $(D(m,T)/T)$ reported in **[1].** The threshold **value** $D(m^*, T)/T$, corresponding to **the** actual number of manually determined segments m^*, is obtained for each sentence in a training set of50 sentences. An average of the average distortion per frame D^* **was** obtained over this training set; this threshold is used for terminating the DP-recursion when $D(m,T)/T$ (3) falls below D^* for **new** (test) sentences.

Segmentation evaluation

We measure the performance of the segmentation techniques in terms of percent match (%M), percent insertions (%I) and percent deletions (%D). %M gives the percentage of segments, obtained by the automatic segmentation, within a specified interval of M frames (or $\delta = 20M$ ms, corresponding to a 20ms analysis frame and 20 ms frame shin) of the manual segments. %I gives the percentage of segments obtained by the automatic segmentation without a corresponding manual segment within the interval of δ ms. %D gives the percentage ofmanually obtained segments without any corresponding automatically segmented boundary within the interval of δ ms. Note that %M + %D = 100; %I can be greater than 100 for high aver-segmentation

Figs. I and 2 show the (%M, %I, %D) for 1 frame tolerance (δ = 20 ms interval), for different parametric representations for the two segmentation techniques, STM and ML segmentation, respectively. These results are for clean speech (36 dB) and for noisy speech with SNRs of -10 dB, -5 dB, 0 dB, 5 dB and 10 dB. In the ML segmentation (Fig. 2), feature set LPC is used along with the likelihood ratio (LR) distortion (8) and all other feature sets use the L_2 distance (unweighted Euclidean distance) (5).

The main points to be noted from the two figures are: i) MFCC-L with ML gives the best performance over other parameters with %M of 85% for clean speech; MFCC-L is also the best in terms of least insertions (%I) and deletions (%D), ii) MFCC with lifter (MFCC-L) for ML segmentation (Fig. 2) gives the best performance over all the SNR's considered, thus being more robust to noise than other parametric representations in terms of (%M, %I, %D), iii) although the ML (8) suggests that LPC with the likelihood ratio (LR) distortion is an optimal solution, MFCC-L can be seen to perform significantly better than LPC with LR for all SNR's considered, iv) LPC with LR distance performs poorer than several other feature sets at all SNR's for both STM and ML segmentation, v) ML segmentation is better than STM on all measures for the best feature set MFCC-L, vi) for STM segmentation, segment match (%M) can be seen to show a marginal increase with additive noise for the parameters CEPS, LAR, RC and MFCC. However, insertions (%I) can also be observed to have increased significantly at lower SNR's. This is because of over-segmentation for noisy speech, which in turn facilitates increased segment match (%M). For bath STM and ML, the threshold parameters for the noisy case are learnt from noisy training data; by this, we are able to limit the over-segmentation. and, vii) several of the other parameters such as CEPS, LAR and RC show similar performance, all being robust to the same extent, particularly for low SNR's of -5 dB and -10 dB; their overall performance falls in between that of LPC and MFCC-L for different SNR's considered.

In Table I, we give the (%M, %I, %D) for ML segmentation with different intra-segment distortions for the case of clean speech. While likelihood-ratio (LR) is applied for LPC, other distances are applied to MFCC-L, the best feature set. L_2 offers the best performance and L_1 and L_3 perform comparably Whereas, the weighted Euclidean (WED) and Mahalanobis (MD) distances perform poorly in comparison to L_2. We would have expected better segmentation performance using WED or MD as they give a more accurate measure of acoustic homogeneity within a segment, by virtue of modeling the vectors within a segment by a multi-variate Gaussian. The $D(m, T)$ (3) for L_2, WED and MD show that this expectation is valid in terms of the distortion associated with the ML segmentation, i.e., WED and MD show five times lower error in terms of the overall distortion obtained, *though* not offering improved segment match with manual segmentation. However, the poorer segmentation performance of WED and MD may be attributed to poor covariance matrix estimation with the limited data in each segment or that the vector elements themselves are noisy. The results of the optimum lifter for MFCC support the latter view. The LR distortion (for LPC) performs poorer than L_2 norm on MFCC-L, as was also shown in Fig. 2.

Table I
(%M, %I, %D) for ML segmentation with different intra-segment distortions: L_1, L_2, L_3, weighted · Euclidean distance (WED), Mahalanobis distance (MD) and likelihood ratio (LR)

	%M	%I	%D	$D(m,T)$
L_1	82.5	19.5	17.5	1646.0
L_2	84.7	16.3	15.3	3206.0
L_3	81.2	20.7	18.8	5324.0
WED	82.4	15.2	17.6	646.2
MD	80.9	17.7	19.1	557.9
LR	71.8	30.1	28.2	3394.0

5. CONCLUSIONS

We have studied in detail the segmentation performance of two segmentation techniques, i) the spectral transition measure (STM) based segmentation and, ii) the maximum likelihood (ML) segmentation for different parametric representations, distance measures and weighting coefficients. For ML segmentation. we have considered intra-segment distortion measures based on likelihood-ratio distortion for the LPC parameten and the Euclidean distance for other parameten. We have conducted a detailed study on the robustness of the different parametric representations to additive white noise. We find that MFCC features along with a new liner given by $1 + 4\sin^{1/2}(\pi n/L)$, $0 \leq n \leq L - 1$, for MFCC dimension L, provides the best performance compared to all LPC derived parameters in both clean and noisy speech upto -10 dB SNR.

6. REFERENCES

[1] M. Bacchiani and M. Ostendorf. Joint lexicon, acoustic unit inventory and model design. *Speech Communication*, 29:99–114, 1999.

[2] L. R. Rabiner and B.-H. Juang. *Fundamentals of speech recognition*. Prentice Hall, 1993.

[3] T. Svendsen and F. K. Soong. On the automatic segmentation of speech signals. In Proceedings *of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 77–80, Dallas, 1987.

[4] E. Vidal and A. Marzal. A review and new approaches for automatic segmentation of speech signals. In L. Torres, E. Masgra, and M. A. Lagunas, editors, *Signal Processing ? Theories and Applications*, pages 43–53. Elsevier Science Publishers B.V., 1990.