A DURATION–BASED CONFIDENCE MEASURE FOR AUTOMATIC SEGMENTATION OF NOISE CORRUPTED SPEECH *

Bryan L. Pellom and *John H.L. Hansen*
Robust Speech Processing Laboratory
Duke University, Box 90291, Durham, NC 27708-0291
http://www.ee.duke.edu/Research/Speech bp@ee.duke.edu jhlh@ee.duke.edu

ABSTRACT

In this study, a duration–based measure is formulated for assigning confidence scores to phonetic time–alignments produced by an automatic speech segmentation system. For speech corrupted by additive noise or telephone channel environments, the proposed confidence measure is shown to provide a reliable means by which gross segmentation errors can be automatically detected and marked for human hand correction. The measure is evaluated by computing Receiver Operating Characteristic (ROC) curves to illustrate the expected trade–off in probability of detecting gross segmentation errors versus false alarm rates.

1. INTRODUCTION

In recent years there has been increased interest in transitioning speech technologies from laboratory settings into real–world environments. Consequently, the demand is high for (1) new algorithms which mitigate the effects of environmental noise, and (2) carefully collected development and evaluation speech corpora recorded in realistic environments. In order to provide phonetically labeled databases, numerous methods for automatic segmentation have been proposed for noise–free environments [1, 2, 3, 4]. In noisy environments, segmentation accuracy drops considerably and the degree to which human experts must hand–correct misplaced phonetic boundaries is dramatically increased [5, 6, 7]. For example, in [7], it was shown that for a Hidden Markov Model (HMM) based segmentation system, 86% of phoneme boundaries are placed within 20 msec of hand–labeled locations when time–aligning noise–free speech. When computer fan noise is added at an SNR of 10 dB, the performance is reduced to 63% within 20 msec if noise compensation is not performed. In the same study, compensation methods including speech enhancement and model adaptation were considered improve noise robustness. Although compensation methods were shown to improve time–alignment accuracy, there remains a sizable performance gap between a compensated system and the same system retrained from hand–labeled speech recorded in the matched noisy environment. Consequently, this paper considers methods of assigning levels of confidence to automatically derived phonetic time–alignments so that the efforts of manual corrections can be better focused.

*This work was supported in part by a National Science Foundation Graduate Research Fellowship and by the U.S. Government.

2. ALGORITHM FORMULATION

2.1. Impact of Noise on Segmentation

Automatic segmentation of speech is a difficult task even in noise–free environments. Often, the "correct" placement of phonetic boundaries is highly subjective, especially during continuous events such as vowel–to–vowel or vowel–to–semivowel transitions. In noise, the boundaries between phones become less sharp and some events are lost entirely (e.g., fricatives spoken in the presence of wideband noise distortion). Fig. 1 illustrates these common problems. For example, in Fig. 1A, the spectrogram for a (8 kHz sampled) phrase from the TIMIT database, "with mock distaste" is shown with the corresponding hand–labeled phonetic transcription. In Fig. 1B, the speech is corrupted by additive computer cooling fan noise (5 dB SNR). Time–alignments were generated by first compensating the HMM–based system with the Parallel Model Combination (PMC) technique [13]. By comparing the two phonetic time–alignments it is clear that there are locations in which the accuracy of the boundary placement can be very poor. Moreover, the resulting durations of some phonetic segments are increased beyond what would normally be expected in natural speech, while others are severely compressed.

2.2. Confidence Measure Formulation

There has been considerable interest in recent years towards improving speech recognition by associating levels of confidence to automatically recognized words [11, 12]. In general, these methods assign a score, $C(w_i)$, for the ith recognized word w_i such that $C(w_i) = 1$ if the word has been correctly recognized and $C(w_i) = 0$ otherwise. Successful measures for speech recognition should adequately predict when recognition fails (i.e., the measure should correlate highly with the actual system performance). Assigning similar confidence scores to time–aligned speech poses several new challenges. First, there exists a continuum of possible boundary misalignments rather than a binary "correct/incorrect" recognition decision. Second, some errors are more significant than others (e.g., consider misalignments involving stop–to–vowel compared with vowel–to–vowel transitions). Ideally the confidence measure should provide user feedback on the quality of the boundary assigned to each phonetic transition. However, significant misalignment of one phoneme will result in misalignments for neighboring phonemes. Therefore, determining the exact location of the initial error is difficult.

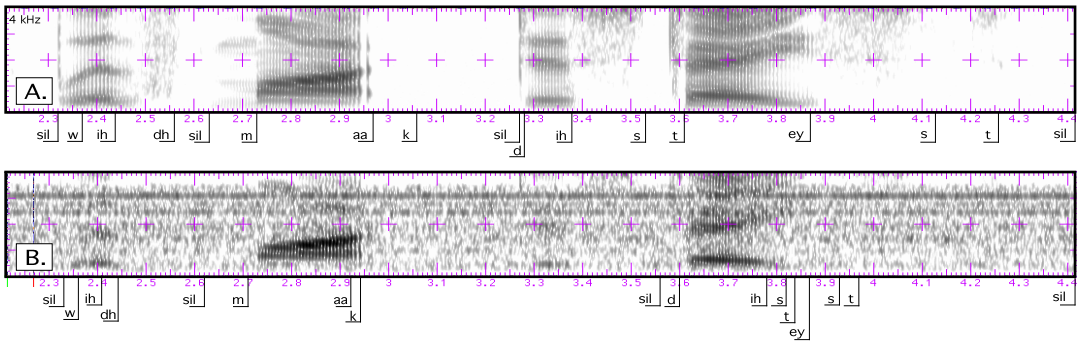

Figure 1: Illustration of common problems encountered by an automatic segmentation system when time–aligning noise corrupted speech. In (A) a spectrogram of the phrase "with mock distaste" is shown with hand–labeled phonetic boundaries. In (B), the speech was degraded by additive computer fan noise (5 dB SNR) and automatically time–aligned by an HMM–segmentation system which was compensated using Parallel Model Combination (PMC).

In addition, the measure should be relatively insensitive to noise or recording conditions. Acoustic features such as log–likelihood scores from HMMs are sensitive to mismatch between training and testing environments. Furthermore, measures of spectral variation have been shown to be sensitive to noise [5]. Consequently, this paper introduces a duration–based confidence measure which is assigned to each sentence rather than on a per–phone basis.

Severe segmentation errors produce phonemes whose durations deviate significantly from that of natural speech. Therefore, to characterize this situation, we first assume that natural phone durations are modeled using a 2–parameter Gamma pdf [10],

$$\mathrm{P}(d_n \mid \ell_n, \alpha, \beta) = \frac{\beta^{-\alpha} d_n^{\alpha-1}}{\Gamma(\alpha)} \exp\left(\frac{-d_n}{\beta}\right) \tag{1}$$

where d_n is the duration variable (in msec), α and β represent parameters of the Gamma pdf for the nth phoneme ℓ_n. Certainly, the duration densities will be dependent on their surrounding phonetic context. In this paper, a set of 18 left–context classes and 19 right–context classes described in [2] are used in estimating the Gamma pdfs.

The observed phoneme duration, d_{obs}, as automatically segmented by a computer algorithm can be modeled as,

$$d_{obs} = d_{act} + e_l + e_r, \tag{2}$$

where d_{act} represents the actual underlying phoneme duration and e_l and e_r represent error made by the system in placing the left and right phonetic boundaries. Here, d_{act} is modeled by a Gamma distribution while each of the error terms are assumed to be statistically independent and modeled by a zero–mean Gaussian distribution, $\mathcal{N}(0, \sigma_e^2)$. The total duration error is therefore given by, $\{\mathcal{E} = e_l + e_r\}$, which also has zero–mean and variance given by, $\{\sigma_{\mathcal{E}}^2 = \sigma_e^2 + \sigma_e^2 = 2\sigma_e^2\}$. Therefore,

$$\mathrm{P}(\mathcal{E}) = \frac{1}{\sqrt{4\pi\sigma_e^2}} \exp\left(\frac{-\mathcal{E}^2}{4\sigma_e^2}\right) \tag{3}$$

The confidence measure is formulated by posing a two–hypothesis problem. First, a duration error threshold, τ, is selected such that errors greater than τ msec are considered unacceptable while errors in duration of less than τ msec are considered acceptable. Since it is generally accepted that phoneme boundary misalignments of more than 20 msec are considered gross errors, in this paper $\tau = 20$ msec. Hypothesis $\mathcal{H}1$ models the case that d_{obs} resulted due to a duration error of $(|\mathcal{E}| > \tau)$ msec. Under $\mathcal{H}0$, we hypothesize that d_{obs} resulted due to a duration error of $(|\mathcal{E}| \leq \tau)$ msec. In other words, hypothesis $\mathcal{H}1$ models the condition that a significant alignment error has occurred while $\mathcal{H}0$ models less severe misalignments. Formally, the detector is given by the likelihood ratio,

$$\lambda(d_{obs}) = \frac{\mathrm{P}(d_{obs} \mid \mathcal{H}1 : \ell, \alpha, \beta, |\mathcal{E}| > \tau)}{\mathrm{P}(d_{obs} \mid \mathcal{H}0 : \ell, \alpha, \beta, |\mathcal{E}| \leq \tau)}. \tag{4}$$

where,

$$\mathrm{P}(d_{obs} \mid \mathcal{H}1 : \ell, \alpha, \beta, |\mathcal{E}| > \tau) = \int_{|\mathcal{E}|>\tau} \mathrm{P}(d_{act} = d_{obs} - \mathcal{E} \mid \ell, \alpha, \beta)\mathrm{P}(\mathcal{E})d\mathcal{E} \tag{5}$$

$$\mathrm{P}(d_{obs} \mid \mathcal{H}0 : \ell, \alpha, \beta, |\mathcal{E}| \leq \tau) = \int_{|\mathcal{E}|\leq\tau} \mathrm{P}(d_{act} = d_{obs} - \mathcal{E} \mid \ell, \alpha, \beta)\mathrm{P}(\mathcal{E})d\mathcal{E}. \tag{6}$$

Substituting (5) and (6) into (4) yields,

$$\lambda(d_{obs}) = \frac{\int_{|\mathcal{E}|>\tau} \mathrm{P}(d_{act} = d_{obs} - \mathcal{E} \mid \ell, \alpha, \beta)\mathrm{P}(\mathcal{E})d\mathcal{E}}{\int_{|\mathcal{E}|\leq\tau} \mathrm{P}(d_{act} = d_{obs} - \mathcal{E} \mid \ell, \alpha, \beta)\mathrm{P}(\mathcal{E})d\mathcal{E}}. \tag{7}$$

Finally, substituting (1) and (3) into (7) and simplifying gives,

$$\lambda(d_{obs}) = \frac{\int_{|\mathcal{E}|>\tau} (d_{obs}-\mathcal{E})^{\alpha-1} \exp\left(\frac{-(d_{obs}-\mathcal{E})}{\beta} - \frac{\mathcal{E}^2}{4\sigma_e^2}\right) d\mathcal{E}}{\int_{|\mathcal{E}|\leq\tau} (d_{obs}-\mathcal{E})^{\alpha-1} \exp\left(\frac{-(d_{obs}-\mathcal{E})}{\beta} - \frac{\mathcal{E}^2}{4\sigma_e^2}\right) d\mathcal{E}}, \tag{8}$$

where the numerator and denominator terms are evaluated with the additional constraint that : $(d_{obs} - \mathcal{E}) > 0$.

Assuming each of the observed durations are statistically independent, an overall log–likelihood score can be computed for a sequence of N phonemes $L = \{\ell_1, \ell_2, \ldots, \ell_N\}$

with observed durations $D = \{d_1, d_2, \ldots, d_N\}$. The proposed confidence measure is computed for each parsed sentence by averaging the log–likelihoods over the entire phone sequence. Specifically,

$$C(D, L) = \frac{1}{N} \sum_{n=1}^{N} \log \lambda(d_n). \quad (9)$$

Intuitively, $C(D, L)$ will be small in value when the observed phone durations deviate little from that expected from hand–labeled speech. Thus, the proposed confidence measure is compared with a threshold, Θ, in order to decide if the segmented sentence requires hand–correction, $\{C(D, L) \geq \Theta\}$, or is acceptable, $\{C(D, L) < \Theta\}$.

3. ALGORITHM EVALUATION

3.1. Baseline Segmentation Algorithm

The segmentation algorithm used in this study was previously formulated in [7]. Each of 46 phoneme units are modeled using a 5–state left–to–right continuous density HMM. For each state, 16 mixture densities are used to characterize the observation pdf. Here, observation vectors consisting of 12 MFCC, 12 delta MFCC, and normalized log–frame energy are computed every 5 msec. The baseline algorithm was evaluated by forced alignment of the complete test set of the 8 kHz resampled TIMIT database (1344 sentences). Performance was determined by computing the absolute distance (in msec) between the automatically determined and hand–labeled phone boundaries. The baseline segmentation accuracy was found to be: $\{47.9\% < 5\text{msec}\}$, $\{69.9\% < 10\text{msec}\}$, $\{85.9\% < 20\text{msec}\}$, $\{95.9\% < 40\text{msec}\}$, $\{98.4\% < 60\text{msec}\}$. These results are comparable with previous systems reported in [2, 3].

3.2. Confidence Measure Evaluation

Speech data from four non–ideal environments were segmented using the baseline system. These included : (8 kHz sampled) TIMIT degraded by additive computer fan noise and additive car noise (5dB SNR), NTIMIT telephone database [8], and CTIMIT cellular database [9]. The PMC technique [13] was used to compensate the baseline system for the additive noise environments while Cepstral Mean Normalization (CMN) was used for the two telephone environments. The confidence measure was computed for each segmented sentence. Fig. 2(A,C,E,G) illustrates scatter plots of the measure's output versus maximum phoneme boundary misalignment for each segmented sentence. The measure was found to range in value from 0 to 6 with the largest output for sentences which contained at least one severe boundary misalignment. This is not surprising given that the numerator term in (8) is large when phone durations deviate significantly from the duration distributions of the hand–labeled training data.

By varying the decision threshold, Θ, a trade–off in detecting a segmented sentence containing a severe misalignment versus false alarm probability can be determined. In Fig. 2(B,D,F,H) ROC curves are shown for the case of detecting a sentence containing a boundary misalignment of at least 100, 300, or 500 msec[1]. For example, 60% of NTIMIT sentences containing a misalignment of at least 300 msec were detected with a false alarm rate of only 10%.

Once a desired operating point has been selected, the confidence measure can be used to automatically alert the user if a segmented sentence requires hand–correction. Table 1 illustrates system performance before and after environmental compensation. For the CTIMIT database, 46.6% of the labeled boundaries are within 60 msec if no compensation is performed. With CMN compensation, the performance improves to 79.5% within 60 msec. The decision threshold for each environment was set such that the false alarm probability was 0.1 for the case of detecting a segmentation error of at least 100 msec. Next, sentences were marked as either "acceptable" $\{C(D, L) < \Theta\}$ or "unacceptable" $\{C(D, L) \geq \Theta\}$ and the alignment performance for each condition was determined. For example, the "unacceptable" CTIMIT sentences had 66.2% of the phoneme boundaries placed with 60 msec compared with 93.7% for those marked as "acceptable". Similar results were obtained for each of the three remaining environments.

There are several additional uses of the proposed confidence measure. For example, one could initially segment a noisy speech corpus using the noise/channel compensated algorithm. Then, using the confidence measure as a guide, parsed phones from sentences with high–confidence (i.e., high–quality) can be used to retrain the system for the noisy environment. During a second pass, the retrained system can then be used to obtain an improved time–alignment of the entire corpus. Finally, on a third pass, sentences marked as requiring hand–correction can then be further examined.

4. CONCLUSIONS

In this paper a new duration–based confidence measure was formulated for automatic segmentation of speech recorded in non–ideal environments. The proposed confidence measure exploits the fact that, in noisy channel–corrupted environments, poor time–alignments result in phones whose durations deviate significantly from that expected of natural speech. The duration–based confidence measure is compared to a threshold which is used to provide user–feedback in an integrated automatic speech time–alignment tool. The measure was tested in 2 additive noise and 2 telephone channel environments and shown to successfully separate low–quality from higher–quality phonetic transcriptions.

5. REFERENCES

[1] T. Svendsen, F.K. Soong, "On the Automatic Segmentation of Speech Signals," *ICASSP*, pp. 77-80, 1987.

[2] A. Ljolje, M.D. Riley, "Automatic Segmentation and Labeling of Speech," *ICASSP*, pp. 473-476, 1991.

[3] F. Brugnara, D. Falavigna, M. Omologo, "Automatic Segmentation and Labeling of Speech based on Hidden Markov Models," *Speech Comm.*, Vol. 12, pp. 357-370, 1993.

[1] For additive car noise there is insufficient data points to plot ROC curves for 300 and 500 msec alignment error detection.

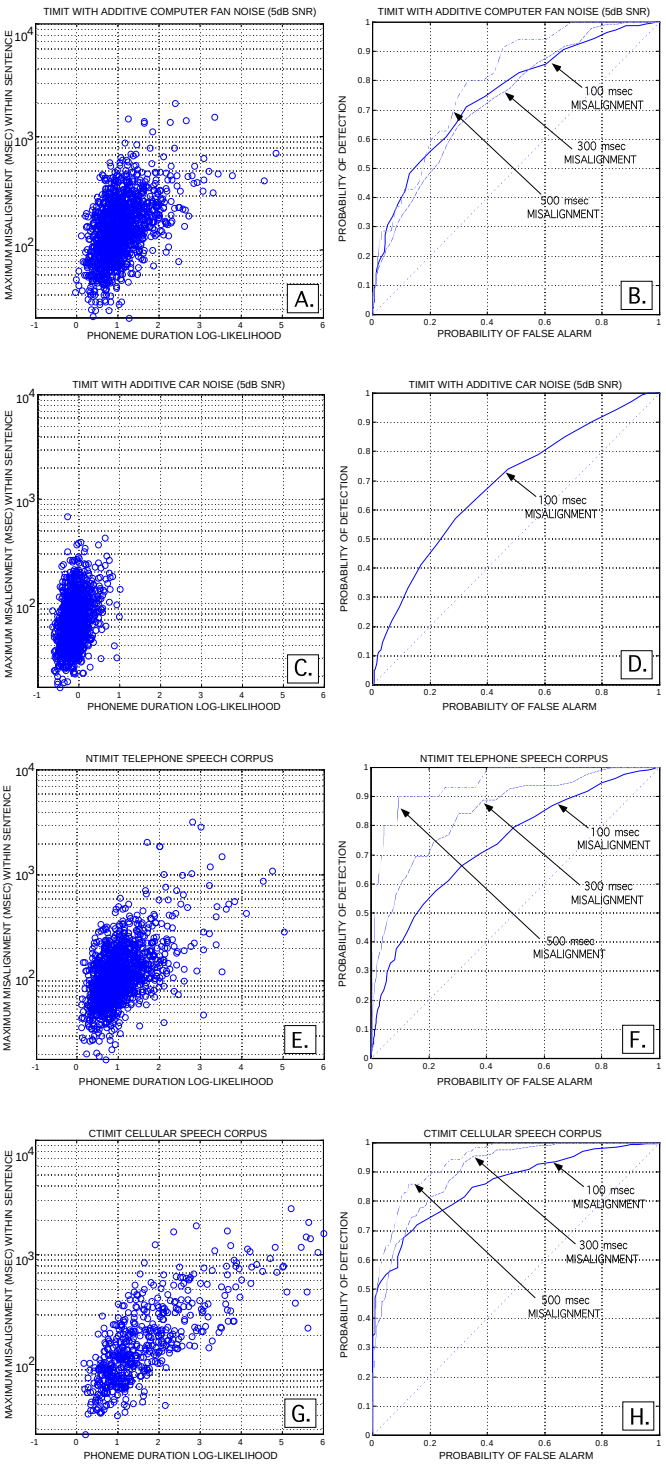

Figure 2: Phoneme duration likelihood score $C(D, L)$ versus maximum within sentence phoneme boundary misalignment (shown in A,C,E,G). In (B,D,F,H), probability of detecting a gross segmentation error versus false alarm probability is shown for 3 phoneme boundary tolerances. Results are shown for detectability of 100, 300, and 500 msec boundary misalignment. Noise conditions are shown for TIMIT sentences degraded by Computer Fan Noise at a 5dB SNR (A,B), TIMIT sentences degraded by Automobile Highway noise at a 5dB SNR (C,D), NTIMIT telephone database (E,F), and CTIMIT cellular database (G,H).

	total # sent.	Boundary Misalignment (msec) ≤ 5	≤ 10	≤ 20	≤ 40	≤ 60
(A) TIMIT + Computer Cooling Fan Noise (5 dB SNR)						
baseline	1344	17.8	26.5	36.2	45.8	52.6
baseline, PMC	1344	32.7	50.1	66.3	79.2	85.7
$C(D,L) < \Theta$	926	35.0	53.6	70.6	83.6	89.6
$C(D,L) \geq \Theta$	418	27.2	41.6	55.9	68.7	76.1
(B) TIMIT + Automobile Highway Noise (5 dB SNR)						
baseline	1344	35.8	53.2	69.0	82.4	88.8
baseline, PMC	1344	43.7	64.7	81.8	93.7	97.2
$C(D,L) < \Theta$	1171	44.3	65.6	82.7	94.3	97.6
$C(D,L) \geq \Theta$	173	39.0	57.4	74.6	88.4	93.8
(C) NTIMIT Telphone Speech Corpus						
baseline	1344	25.1	40.3	56.0	67.3	72.8
baseline, CMN	1344	32.1	52.3	72.8	82.3	88.6
$C(D,L) < \Theta$	1010	33.2	54.4	75.8	88.8	93.7
$C(D,L) \geq \Theta$	334	28.4	45.2	62.8	74.8	80.9
(D) CTIMIT Cellular Telephone Speech Corpus						
baseline	548	17.7	26.0	34.6	42.6	46.6
baseline, CMN	548	29.7	45.4	60.4	73.4	79.5
$C(D,L) < \Theta$	274	36.1	55.3	73.1	87.1	92.6
$C(D,L) \geq \Theta$	274	23.2	35.3	47.5	59.4	66.2

Table 1: Automatic segmentation accuracy for baseline HMM time–alignment, noise/channel compensated HMM time–alignment, and performance for sentences with confidence scores below and above a confidence threshold of Θ.

[4] A. Vorstermans, J.-P. Martens, B. Van Coile, "Automatic Segmentation and Labelling of Multi-Lingual Speech Data," *Speech Comm.*, Vol. 19, pp. 271-293, 1996.

[5] M.W. Mak, W.G. Allen, "Spectral Transitivity Functions for Speech Segmentation in Noise," *Acoustic Letters*, Vol. 16, No. 10, pp. 228-234, 1993.

[6] B. Petek, O. Andersen, P. Dalsgaard, "On the Robust Automatic Segmentation of Spontaneous Speech," *ICSLP*, pp. 913-916, 1996.

[7] B. L. Pellom, J. H.L. Hansen, "Automatic Segmentation and Labeling of Speech Recorded in Unknown Noisy Channel Environments," *Speech Communication: Special Issue on Robust Speech Recognition*, (to appear, Nov. 1998).

[8] C. Jankowski, A. Kalyanswamy, S. Basson, J. Spitz, "NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database," *ICASSP*, pp. 109-112, 1990.

[9] K. Brown, E. George, "CTIMIT: A Speech Corpus for the Cellular Environment with Applications to Automatic Speech Recognition," *ICASSP*, pp. 105-108, 1995.

[10] S. E. Levinson, "Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition," *Comp. Speech and Lang.*, Vol. 1, pp. 29-45, 1986.

[11] M. Siu, H. Gish, F. Richardson, "Improved Estimation, Evaluation, and Applications of Confidence Measures for Speech Recognition," *Eurospeech*, Vol. 2, pp. 831-834, 1997.

[12] L. Gillick, Y. Ito, J. Young, "A Probabilistic Approach to Confidence Estimation and Evaluation," *ICASSP*, Vol. 2, pp. 879-882, 1997.

[13] M. J.F. Gales, S. J. Young, "Cepstral Parameter Compensation for HMM Recognition in Noise," *Speech Comm.*, Vol. 12, pp. 231-240, 1993.