

# Speech Rate Normalization Used to Improve Speaker Verification

C.J. van Heerden and E. Barnard

Human Language Technology Group, Meraka Institute, CSIR / University of Pretoria

cvheerden@csir.co.za, ebarnard@csir.co.za

## Abstract

A novel approach to speech rate normalization is presented. Models are constructed to model the way in which speech-rate variation of a specific speaker influences the duration of phonemes. The models are evaluated in two ways. Firstly, the mean square error in phoneme duration based on our normalization is compared to the same error when such normalization is not applied. The second evaluation uses the durations of context-dependent phonemes in speaker verification. Both methods show that this approach to normalization is indeed effective to counteract the effect of variable speaking rates.

## 1. Introduction

Speech is the most natural way for humans to communicate with each other. Over the past decade, much work has been done in man-machine communications in order to incorporate speech as a new modality in multimedia applications [1]. We are interested in two particular disciplines which have received considerable interest: speech recognition, in which the aim is for the machine to extract and understand the linguistic message in the speech, and speaker recognition, where the goal is to identify, recognize or verify the speaker responsible for producing the speech. There are several factors that have limited the success of integrating speech into machine communications such as transmission line degradations, channel mismatch and speech rate variability [2]. Speech rate variability has been found to be significant in increasing the error rate of speech recognition, especially when it deviates greatly from the training data [2]. Several ways have been proposed to remove this speech rate variability by speech rate normalization, with some recent proposal such as [3].

We have previously shown that phoneme durations are high-level features that can be used effectively in speaker verification [4]; here, we confirm those findings on a larger corpus, and demonstrate that speech-rate normalization can be applied to further improve the accuracy of this feature. The duration model is described in detail, in particular the way in which it was constructed. The model is then applied in a novel speech rate normalization technique on the YOHO corpus and the resulting normalized durations are submitted to a speaker verification system. The equal error rate (EER) using the normalized durations is then compared with the EER using unnormalized durations, and also with the EER when duration information is not employed.

## 2. Proposed phoneme duration modeling

### 2.1. Choosing parametric models

Since the duration of a phoneme is known to depend on its acoustic context, we model the durations of context-dependent

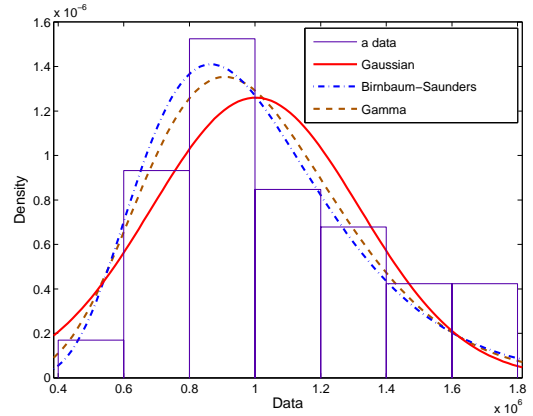


Figure 1: Distribution of durations of the triphone “s-eh+v” for a single speaker with different distribution functions fitted to it.

phonemes (from here on referred to as triphones). These durations are obtained by forced alignment of each YOHO utterance, using the known transcription and the speaker-specific acoustic model described above. Only one pronunciation per word was allowed, thus resulting in 49 triphones. To decide which parametric model to use for the duration density functions of the triphones, several parametric forms were fit to the triphone durations obtained in this fashion. Typical results for the speaker-specific and speaker-independent distributions are shown in Figures 1 and 2, respectively. The histogram density estimates, as shown by the bar graphs in these figures, are consistently unimodal, suggesting that a single parametric component will be sufficient. For the speaker-specific density function of Figure 1, the Birnbaum-Saunders and Gamma distributions seem to fit the data best, whereas the normal distribution provides a better fit for the speaker-independent case (Figure 2). However, all these differences are fairly small, and we have therefore used a normal distribution in all the experiments described below.

### 2.2. Details of triphone duration models

The models used in our tests were constructed for each triphone  $k$  by calculating the sample mean

$$\bar{x} = \frac{1}{M} \sum_{n=1}^M x_n, \quad (1)$$

where  $M$  is the number of observations of the triphone and  $x_n$  is the duration of the  $n$ 'th observation. An unbiased estimate of

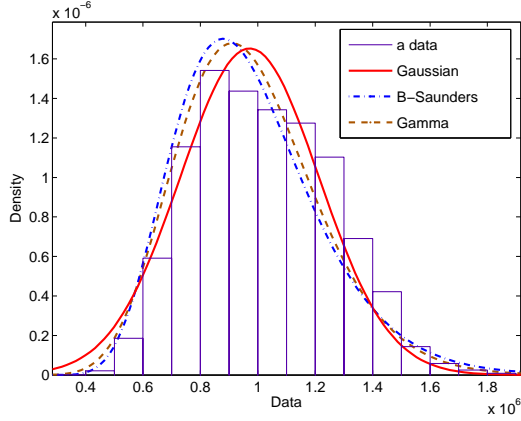


Figure 2: Distribution of durations of the triphone “s-eh+v” for 168 speakers with different distribution functions fitted to it.

the sample variance  $\sigma^2$  was also calculated as

$$s_{M-1}^2 = \frac{1}{M-1} \sum_{n=1}^M (x_n - \bar{x})^2. \quad (2)$$

Every speaker thus has 49 time models of the form  $(\mu, \sigma^2)$ . The time models were constructed by using all the extracted time durations from the 4 enrollment sessions. Testing was then performed by first extracting the time durations of the triphones in the test session and then calculating a score

$$P(x|\bar{x}, s_{M-1}^2) = \frac{1}{\sqrt{2\pi s_{M-1}^2}} e^{-\frac{(x-\bar{x})^2}{2s_{M-1}^2}}, \quad (3)$$

where  $x$  is the observed duration of a specific triphone. The evaluation of (3) yields a value that occurs on the normal distribution with parameters  $(\bar{x}, s_{M-1}^2)$ . This value is normalized by evaluating the normal distribution with the same value, but using the UBM parameters (which are the means and variances of the appropriate context-dependent triphone, calculated across all training sessions by all speakers). A score is then generated for a speaker  $i$  as

$$Score_i = \frac{1}{L} \sum_{l=1}^L \log(P(x_l|\lambda_c)) - \frac{1}{L} \sum_{l=1}^L \log(P(x_l|\lambda_{UBM})) \quad (4)$$

where  $L$  is the number of observed triphones in the test session. Tests were again performed on a rotating scheme as before, where one speaker is the claimed “client” and all speakers excluding the (acoustic) cohort set are tested using the claimed speaker’s models. Once all scores have been obtained, they were again put in an ordered list and the EER was determined.

Figures 3 and 4 illustrate typical distributions of durations observed in our tests. Figure 3 shows an example of a triphone that provides good discrimination between two speakers; in other words, time durations of speaker  $x$  matched to the model of  $y$  would produce a poor score and the general UBM would be chosen, resulting in a correct reject decision of the impostor. Figure 4 illustrates a bad example of a triphone to use, since speaker  $x$ ’s time durations would match speaker  $y$ ’s durations well, resulting in a good score for an impostor.

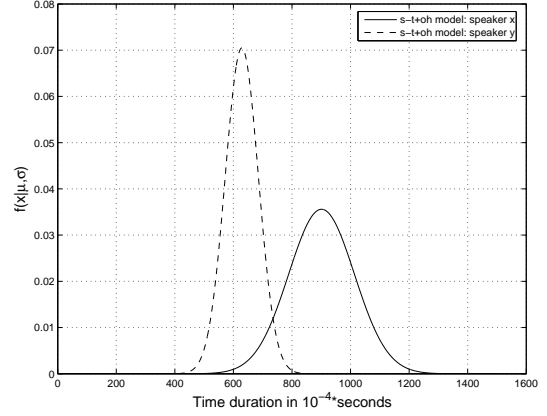


Figure 3: Probability distributions of a triphone that provides good discrimination between a pair of speakers.

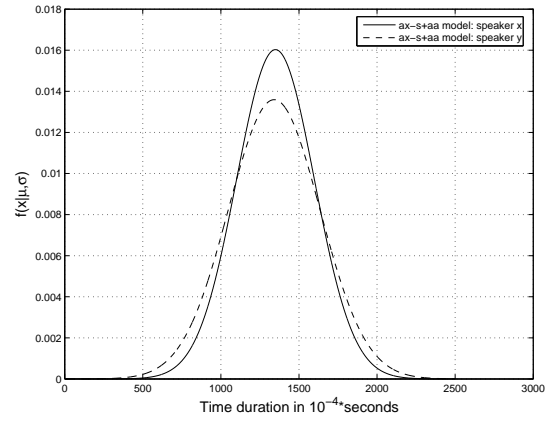


Figure 4: Probability distributions of a triphone that does not provide good discrimination between a pair of speakers.

Since both cases are observed in our data, we used an empirical methodology to determine whether durations are useful for the task of speaker verification.

### 3. Refinement of predicted phoneme duration

In the preceding sections we assumed that the duration of a particular phoneme spoken by a given speaker is described by a normal distribution, independently of the durations of other phonemes in the utterance. This is clearly not realistic - for example, the speaking rate will tend to influence all the phonemes in an utterance in a correlated manner. It is therefore interesting to ask whether a more detailed duration model can be developed, to account for such influences on phoneme durations. A more complete model could also include factors such as the position of the phoneme in the word or utterance, but for now we have concentrated on the influence of speaking rate.

To do this, we developed a model for predicting the duration of a phoneme of the form

$$t(ms) = [ t_{f,s} \quad \chi_{w,s} ] \cdot \lambda_{\mathbf{w},\mathbf{f},\mathbf{s}}^T, \quad (5)$$

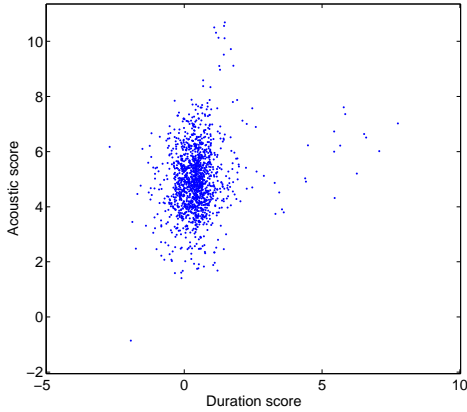


Figure 5: Correlation between the normalized temporal and acoustic features.

where  $t_{f,s}$  is the speaker-specific mean estimate of the phoneme duration for phoneme  $f$  and  $\chi_{w,s}$  is the “stretch factor” for a specific word  $w$  spoken by  $s$ . This is determined as

$$\chi_{w,s} = \frac{\tau - \hat{\tau}}{\sum \sigma_n}. \quad (6)$$

Here  $\tau$  is the true word length,  $\hat{\tau}$  is the estimated word length that was determined by summing the means of the phonemes constituting the word and  $\sum \sigma_n$  is the sum of the standard deviations of these phonemes. Finally,  $\lambda_{w,f,s}$  is the vector of parameters obtained from a General Linear Model (GLM) in order to model the effect of the speech rate on the specific phoneme. The GLM has the form

$$y = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_k X_k \quad (7)$$

and the coefficient vector  $\mathbf{b}$  is determined as

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} \quad (8)$$

This novel speech rate normalization technique was then applied to the testing procedure described above as follows. For every test session, the parameters  $\lambda_{f,s}$  and  $\lambda_{w,s}$  of the claimed speaker were used to normalize the session’s phonemes with regard to speech rate and produce

$$t_{norm} = \frac{t_{measured} \cdot t_{f,s}}{\begin{bmatrix} t_{f,s} & \chi_{w,s} \end{bmatrix} \cdot \lambda_{\mathbf{w},\mathbf{f},s}^{\mathbf{T}}}, \quad (9)$$

### 3.1. The effect of normalization on the correlation with acoustic scores

When two features are fused in order to obtain a new feature, the performance of the feature will only improve if the two features are uncorrelated to a certain degree. High-level features have received considerable interest over the past couple of years and have been shown to contain valuable uncorrelated (to MFCCs) information with regard to speaker identity [5]. It was thus interesting to note that the correlation between our duration features and the MFCCs was only 0.24. After normalizing, the correlation decreases even further to 0.19 - the exact reason for this is an interesting field for further research. Scatterplots of the features can be seen in figures 5 and 6.

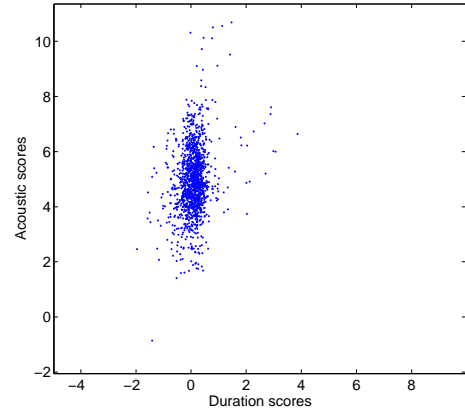


Figure 6: Correlation between the unnormalized temporal and acoustic features.

## 4. Experimental framework: database, protocol and systems

### 4.1. The YOHO corpus

The YOHO corpus is a large supervised speaker verification database [6]. It consists of 138 speakers (106 males and 32 females) who spoke prompted utterances from a restricted grammar set of 56 two-digit numbers ranging from 21-97 [7]. The utterances comprised combination-lock phrases (e.g. 21 – 38 – 44) as proposed by [7]. Four such phrases were prompted during a verification session and 24 such phrases for a training/enrollment session. The YOHO corpus has 4 enrollment sessions per speaker and 10 verification sessions. The data was recorded with a microphone using a 3.8 kHz bandwidth in an office environment with normal background noise. This corpus is widely used for the evaluation of text-dependent systems for speaker verification, and is therefore suitable for the comparative evaluation of such systems.

Table 1: YOHO corpus summary: number of phrases per speaker

Session	Nr of phrases
Enrollment	96
Verification	40
Total	136

### 4.2. Testing procedure

#### 4.2.1. Background

In order to compare the performance of our proposed speaker verification system to that of other speaker verification systems, a standard testing procedure was employed, similar to that used by others on the same corpus (see [1], [8], [9]). The exact test procedure is most clearly described by Reynolds [1] and is described in section (described in section 4.2.2).

Table 2 summarizes the results of several different tests that were performed by Reynolds [1] on the YOHO corpus. (In this table, *msc* denotes “maximally-spread close” and *msf* “maximally-spread far”; these are two different approaches to

selecting cohort speakers – see below.) The test  $M+F(10\text{ msc})$

Table 2: Equal error rates reported in [1] for different experimental conditions.

Test	YOHO(eer)
M(10 msc)	0.20
M(5 msc, 5 msf)	0.28
F(10 msc)	1.88
F(5 msc, 5 msf)	1.57
M+F(10 msc)	0.58
M+F(5 msc, 5msf)	0.51

was used as basis for our comparison, the only difference being that all four enrollment sessions were used for enrolling the speakers. (Reynolds used the fourth session for cohort selection).

In order to perform comparable tests using the temporal features, we had to adapt the use of cohorts for score normalization. A cohort set is a small selection of speakers other than the true speaker, which are used to normalize the speaker’s score. That is, to determine whether the true speaker ( $Pr(\lambda_c|X)$ ) or an impostor ( $Pr(\lambda_{\bar{c}}|X)$ ) is speaking, we compute the likelihood ratio:

$$\text{likelihoodratio} = \frac{Pr(\lambda_c|X)}{Pr(\lambda_{\bar{c}}|X)} \quad (10)$$

In (10)  $X$  denotes the spoken utterance,  $\lambda_c$  the claimed speaker model and  $\lambda_{\bar{c}}$  the cohort (also known as background or impostor) model. By applying Bayes’ rule and discarding the constant prior probabilities for claimant and impostor speakers (they are accounted for in the decision threshold) [1] and working in the log domain, (10) can be rewritten as

$$\Lambda(X) = \log p(X|\lambda_c) - \log p(X|\lambda_{\bar{c}}) \quad (11)$$

The speaker is accepted as the claimed speaker if  $\Lambda(X) > \theta$  and rejected as an impostor if  $\Lambda(X) < \theta$  where  $\theta$  is an appropriate threshold [1].  $\theta$  can be speaker specific (which is computationally more expensive, but also more accurate) or global. The determination of the EER in our test used a global threshold approach, as in [1].

This standard approach to normalization works well if only one type of feature is employed. However, the choice of cohort speakers dictates a group of speakers that cannot be tested as possible impostors, which complicates the procedure when a second feature set is to be used. (If the cohort speakers are based on acoustic features only, they will not necessarily be a good model when using the time feature.) We therefore chose to normalize the temporal features using a universal background model (UBM) rather than a cohort set while still using a UBM approach for the acoustic feature.

#### 4.2.2. Detailed test description

The HTK 3.2.1 toolkit [10] was used to construct the speaker verification system. MFCCs were used as input features together with delta and acceleration coefficients. HMMs with one Gaussian mixture per state were created for all context-dependent triphones occurring in the restricted grammar set.

A cohort set of 10 speakers were selected for every speaker in the database in accordance with the procedure in [1]. Choices that arise with background speakers are the choice of specific speakers and the number of speakers to employ. The selection can be viewed from two different points of view. Firstly, the

background set can be chosen in order to represent impostors that sound similar to the speaker, referred to as dedicated impostors [1]. Another approach is to select a random set of speakers as the background set, thus expecting casual impostors who will try to represent a speaker without consideration of sex or acoustic similarity. By selecting the dedicated impostor background set, in contrast, the system may be vulnerable to speakers who sound very different from the claimed speaker [7].

The selection of the background set was done on a per speaker basis and it was decided to use the dedicated impostor approach [1].

For speaker  $i$ , all other speakers (excluding  $i$ ’s cohort set of 10 speakers) were then used as impostors and tested using (11). Speaker  $i$ ’s verification data was also tested using (11), resulting in 1270 impostor attacks and 10 true attempts to gain access to the system (since every speaker has 10 verification sessions). This process was repeated for all speakers in the corpus, resulting in 175260 impostor attacks and 1380 true attempts.

In particular, (11) was evaluated as follows using the cohort set and the claimed speaker model: First,  $\log p(X|\lambda_c)$  was evaluated as

$$\log p(X|\lambda_c) = \frac{1}{T} \sum_{t=1}^T \log p(x_t|\lambda_c), \quad (12)$$

where  $T$  is the number of frames in the utterance and  $\frac{1}{T}$  is used to normalize the score in order to compensate for different utterance durations.

$\log p(X|\lambda_{\bar{c}})$ , the probability that the utterance was from an impostor was calculated using the claimed speaker’s cohort set as

$$\log p(X|\lambda_{\bar{c}}) = \log \left\{ \frac{1}{B} \sum_{b=1}^B p(X|\lambda_b) \right\}, \quad (13)$$

where  $p(X|\lambda_b)$  was calculated as in (12).

The UBM that was used to normalize the phoneme durations in actual fact consists of 49 independent triphone models. It was constructed by simply calculating the mean and variance for each of the 49 triphones, using all observations in the enrollment set over all speakers.

The EER was then calculated by creating a list of all the likelihood ratios, sorting it and finding the threshold point where the percentage of true speakers below the threshold is equal to the percentage of false speakers above the threshold.

## 5. Results

The duration models were tested on the YOHO corpus. A test was conducted to measure the accuracy of the models compared to two other approaches: (a) the duration of each triphone is assumed to be constant and (b) the duration of each triphone is assumed to scale linearly with the stretch factor. The results are summarized in table 3, and clearly show the importance of rate normalization in accounting for triphone durations. Our general linear model is also seen to be significantly more accurate than a simple linear scaling of phoneme durations.

Speaker verification tests were also conducted with and without the duration normalization procedure, giving the results in table 4. Despite the significant improvement in modelling accuracy achieved with speech-rate normalization, such normalization gives only a small reduction in error rate when duration information is used by itself for speaker verification (7.1% to 6.8%). However, in combination with the acoustic

Table 3: Comparison of three approaches to the modelling of speech rate. The second column contains the mean-squared difference between the actual and predicted phoneme durations (averaged over all test utterances), and third column contains the standard error of this estimate (that is, the standard deviation of all differences divided by the square root of the total number of phonemes in these utterances).

Model	MS error (msec)	Standard error of MS error estimate
Constant speaker-specific duration per phoneme	777.25	65.94
Linear scaling of phoneme durations	522.33	25.46
General linear model of phoneme durations	430.59	16.86

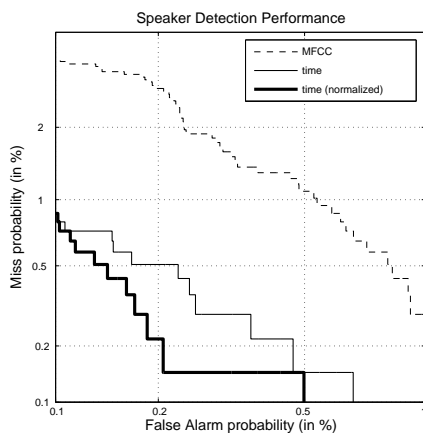


Figure 7: DET curves for MFCC and combined features (time unnormalized and normalized).

scores, the normalization procedure is shown to be highly efficient. This difference is also illustrated in the detection error trade-off (DET) curves shown in Figure 7.

Table 4: Equal error rates obtained on the YOHO database ( $M+F$ , msc).

Feature set	EER before normalization	EER after normalization
MFCCs	0.68%	0.68%
Time	7.1%	6.8%
MFCCs and time	0.29%	0.20%

Our results (0.20%) with the combined acoustic (MFCC) and temporal features are seen to compare favorably with those achieved by other researchers [6]. The temporal features by themselves are significantly less reliable than the acoustic features, but reduce the error rate by a factor of approximately four when combined with those features. The duration features were linearly combined with the acoustic features, the parameters determined by exhaustive testing and optimization. This phenomenon was foreseen by Reynolds et al. [11] when they mentioned that higher-level features need to be investigated and that these would probably not give good performance on their

own (as we experienced), but that they could beneficially be fused with more conventional features to obtain good performance. Similar observations were made in [5] when EERs as high as 26% were observed, but the fusion with conventional features produced an improvement of 71%. Our observations suggest that the duration features are reasonably uncorrelated with the acoustic features, and the scatter plots in figures 5 and 6 confirm this impression.

## 6. Conclusion and future directions

We have shown that phoneme durations constitute a speaker trait that can improve speaker recognition systems. Durations are subject to various influences, such as changes in speaking rate. Tests on the YOHO corpus have confirmed that speech rate normalization can improve the robustness and accuracy of phoneme durations as a feature in speaker recognition. Speech recognition will also benefit from speech rate normalization, as has been proposed by [3]. Further research should be done on other corpora: differing speech rates were not part of the YOHO protocol, and we expect that normalization will be even more significant with more variable corpora.

It was also noted that occasional failures of the automatic alignment process (especially erroneous boundary detection for phonemes at the beginning and end of phrases) contributed significantly to the errors that occur when using temporal information by itself. Rectifying this problem is expected to enhance the power of duration features significantly. A possible remedy for this problem is to incorporate acoustic scores in weighting the duration models. A low acoustic score will then indicate that the particular phoneme has not been recognized with high reliability and can thus be discarded or assigned a lower weight than durations detected with high reliability. This approach can also be used if duration features are to be used in a text-independent application, which will be the next step towards a practical implementation of this feature.

Although this research has been done with an HMM-based text-dependent speaker verification system, results such as those obtained with the text-independent system from [12] suggest that the low correlation observed between temporal and acoustic scores can be beneficial in other classes of speaker verification systems.

The duration models we have used are still rather crude since all triphones are assigned equal weights and are modelled by independent Gaussian distributions. The models can probably be improved by considering other factors such as the frequency of observation of triphones, the acoustic reliability of the observation, correlation between triphones and giving greater weight to more discriminative triphones.

Overall, it seems as if triphone durations are likely to be a useful addition to almost any toolbox for speaker verification system development.

## 7. References

- [1] D.A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, March 1995.
- [2] N. Morgan, E. Fosler, and N. Mirghafori, "Speech recognition using on-line estimation of speaking rate," in *Proceedings of Eurospeech*, September 1997, vol. 4, pp. 2079–2082.
- [3] H.R. Pfitzinger, "Intrinsic phone durations are speaker-

- specific,” in *Proceedings of ICSLP*, September 2002, vol. 1, pp. 1113–1116.
- [4] C.J. van Heerden and E. Barnard, “Using timing information in speaker verification,” in *Proceedings of PRASA*, December 2005, pp. 53–57.
- [5] D.A. Reynolds, W. Andrews, J.P. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and X. Bing, “The SuperSID project: exploiting high-level information for high-accuracy speaker recognition,” in *Proceedings of ICASSP*, April 2003, vol. 4, pp. 784–787.
- [6] J. Campbell, “Testing with the YOHO CD-ROM voice verification corpus,” in *Proceedings of ICASSP*, May 1995, vol. 1, pp. 341–344.
- [7] A. Higgins, L. Bahler, and J. Porter, “Speaker verification using randomized phrase prompting,” *Digital Signal Processing*, vol. 1, no. 2, pp. 89–106, April 1991.
- [8] H-S. Liou and R.J. Mammone, “A subword neural tree network approach to text-dependent speaker verification,” in *Proceedings of ICASSP*, May 1995, vol. 1, pp. 357–360.
- [9] A.E. Rosenberg, J. DeLong, C-H. Lee, B-H. Juang, and F.K. Soong, “The use of cohort normalized scores for speaker verification,” in *Proceedings ICSLP*, October 1992, vol. 1, pp. 599–602.
- [10] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Veltchev, and P. Woodland, *The HTK Book*, Cambridge University Engineering Department, <http://htk.eng.cam.ac.uk/>, 2005.
- [11] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital Signal Processing*, vol. 10, no. 2, pp. 19–41, April 2000.
- [12] L. Ferrer, H. Bratt, V.R.R. Gadde, S. Kajarekar, E. Shriberg, K. Sönmez, A. Stolcke, and A. Venkataraman, “Modeling duration patterns for speaker recognition,” in *Proceedings of Eurospeech*, September 2003, pp. 784–787.