Speaker Recognition from Coded Speech and the Effects of Score Normalization

R.B. Dunn, T.F. Quatieri, D.A. Reynolds, J.P. Campbell

MIT Lincoln Laboratory, Lexington, MA
{rbd,tfq,dar,jpc}@sst.ll.mit.edu

Abstract

We investigate the effect of speech coding on automatic speaker recognition when training and testing conditions are matched and mismatched. Experiments used standard speech coding algorithms (GSM, G.729, G.723, MELP) and a speaker recognition system based on gaussian mixture models adapted from a universal background model. There is little loss in recognition performance for toll quality speech coders and slightly more loss when lower quality speech coders are used. Speaker recognition from coded speech using handset dependent score normalization and test score normalization are examined. Both types of score normalization significantly improve performance, and can eliminate the performance loss that occurs when there is a mismatch between training and testing conditions.

1. Introduction

With the increase in availability and use of digital cellular and VoIP telephony there has been increased interest in the effects of speech compression algorithms on speaker recognition systems. In this paper we investigate the effects of four commonly used speech coding algorithms on automatic speaker recognition for conversational telephone speech. We also examine the effects of a mismatch between the training and testing phases of the speaker recognition system, where, for example, the speaker model is trained from uncoded speech and in the recognition phase the speech is coded.

The speaker recognition experiments in this paper are performed using a Gaussian mixture model universal background model (GMM-UBM) speaker recognition system [1]. This type of speaker recognition system has consistently had excellent performance in the annual NIST Speaker Recognition Evaluations [1, 2]. In the experiments, coded speech is generated by first encoding and then decoding speech segments from the NIST Speaker Recognition Benchmarks [3] (SRB). This simulates the condition where speech originating from a land line is encoded digitally during transmission. This occurs, for example, when a user on a land line is connected to a digital cellular user.

In past work [4] a relatively small data set containing only 50 target speakers per gender was used to examine the effects of speech coding algorithms on the above speaker recognition system. Those experiments demonstrated that there was a slight loss in speaker recognition performance for speech compressed with toll quality coders and that this loss increased slightly for lower quality speech coders. In particular, this performance loss increased when the training and testing conditions were mismatched. These experiments used electret handsets and required that the training and testing speech use the same handset. We show that these results hold for a larger and more challenging data set where carbon button handsets are included and where the handset is allowed to vary between training and testing. These added conditions are known to pose a significantly increased challenge to automatic speaker recognition systems [2]. We also examine the effect of score normalization which is commonly used to boost system performance. We find that handset dependent score normalization (Hnorm) and handset dependent test score normalization (HTnorm) are about at effective at improving system performance for coded speech about as they are for speech that has not been coded. We also find that in most cases score normalization removes the performance loss associated with a mismatch between training and testing conditions.

2. Background

2.1. Speech Coders

We chose four standard speech coding algorithms to cover a range of speech quality and bit rates. The coders are: GSM (12.2 kb/s), G.729 (8 kb/s), G.723 (5.3 kb/s) and MELP (2.4 kb/s). All of the coders use a source/filter representation of speech, with the primary differences being the fidelity of the transmitted parameters and the manner of coding and regenerating the excitation. The GSM coder is the

This work was supported by the Department of Defense under Air Force Contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

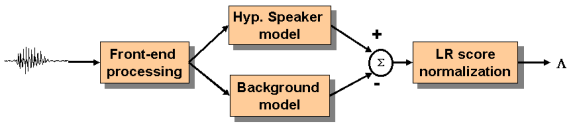

Figure 1: *GMM-UBM likelihood ratio detector.*

ETSI Pan-European standard fixed-point enhanced GSM at 12.2 kb/s and is based on a regular multi-pulse residual coding scheme [5]. The G.729 coder is a fixed point coder at 8 kb/s standardized by ITU-T for personal communication and satellite systems, and is based on a conjugate-structure algebraic CELP residual coding scheme [6]. These two coders produce toll quality speech. The G.723 coder is the CELP-based ITU-T multi-media standard coder at 5.3 kb/s [7]. Finally, the Mixed Excitation Linear Prediction[8] (MELP) coder which is the new U.S. Federal Standard for speech coding at 2.4 kb/s uses a synthetic excitation (harmonics plus noise). This coder is used for narrowband radio and satellite communications.

2.2. Speaker Recognition System

The basic speaker detector is a likelihood ratio detector with target and alternative probability distributions modeled by Gaussian mixture models (GMMs) as shown in Figure 1. A Universal Background Model (UBM) GMM is used as the alternative hypothesis model, and from this, target models are derived using Bayesian adaptation (also known as Maximum A-Posteriori (MAP) training) [1]. The scores are normalized such that a single speaker-independent threshold can be used for detection.

The front end processing for the system is as follows. A 19-dimensional mel-cepstral vector is extracted from the speech signal every 10 ms using a 20 ms window. The mel-cepstral vector is computed using a simulated triangular filterbank on the DFT spectrum. Bandlimiting is then performed by only retaining the filterbank outputs from the frequency range 300-3138 Hz. Cepstral vectors are processed first with cepstral mean subtraction and then with RASTA filtering to mitigate linear channel bias effects. Delta cepstra are then computed over a ± 2 frame span and appended to the cepstra vector producing a 38 dimensional feature vector. Lastly, the feature vector stream is processed through an adaptive, energy-based speech detector to discard low-energy vectors.

The UBM is a 2048 mixture gender-independent, handset-independent GMM trained using about 9 hours of data selected from the 1999 NIST SRB to be approximately evenly divided between sex and handset type. Target models are derived by Bayesian adaptation (a.k.a. MAP estimation) of the UBM parameters using the two minutes of training data. Only the mean vectors are adapted as this has been observed to provide better performance. The amount of adaptation of each mixture mean is data dependent. Details of the adapted GMM-UBM system can be found in [1].

2.3. Score Normalization

In past work [1, 9], the use of score normalization has significantly improved the performance of speaker recognition systems. In this work we undertake to determine if score normalization, in particular handset dependent score normalization (Hnorm) and handset dependent test-score normalization (HTnorm), can be used to improve performance when the speech has been coded. In Hnorm, scores from a handset-dependent collection of fixed non-target (imposter) *speech samples* are used to normalize a speaker model, while in HTnorm, scores from a handset-dependent collection of fixed non-target (imposter) *speaker models* are used to normalize a speech test segment.

In the application of Hnorm, we first compute the log-likelihood ratio scores for a target speaker with a set of imposter test segments coming from both carbon-button (CARB) and electret (ELEC) handsets. We assume these scores have a Gaussian distribution and we estimate the handset-dependent means and standard deviations for these scores. To avoid bimodal distributions, the non-speaker data is of the same gender as the target speaker. The target speaker now has two sets of parameters describing his/her model's response to CARB and ELEC type speech:

$$\{\mu(\text{CARB}), \sigma(\text{CARB}), \mu(\text{ELEC}), \sigma(\text{ELEC})\}$$

In this paper we used 200 30-second speech segments per handset type, per gender derived from the 1999 NIST SRB test corpus. In general, the duration of the speech segments used to estimate Hnorm parameters should match the expected duration of the test speech segments.

During recognition, a handset detector is used to supply the handset type of the test segment. This detector is a simple maximum likelihood detector with handset types represented by 256 mixture GMMs [10]. Then for each test segment, X, Hnorm is applied to the log-likelihood ratio score as

$$\Lambda^{\text{Hnorm}}(X) = \frac{\Lambda(X) - \mu(HS(X))}{\sigma(HS(X))}, \qquad (1)$$

where $HS(X)$ is the handset label for X.

The desired effect of Hnorm is illustrated in Figure 2. This figure shows Log-Likelihood Ratio (LLR) score distributions for two speakers before (left column) and after (right column) Hnorm has been applied. The effect

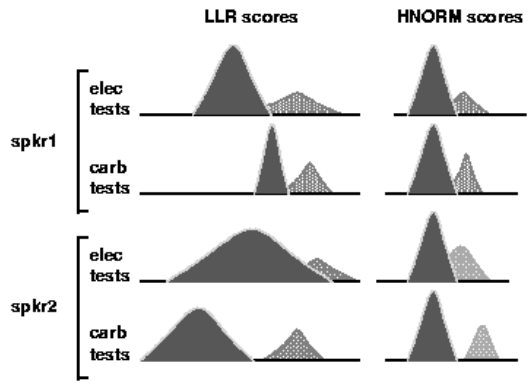

Figure 2: *Pictorial example of Hnorm compensation. This picture shows Log-Likelihood Ratio (LLR) score distributions for two speakers before (left column) and after (right column) Hnorm has been applied. After Hnorm, the non-speaker score distribution for each handset type has been normalized to zero mean and unit standard deviation.*

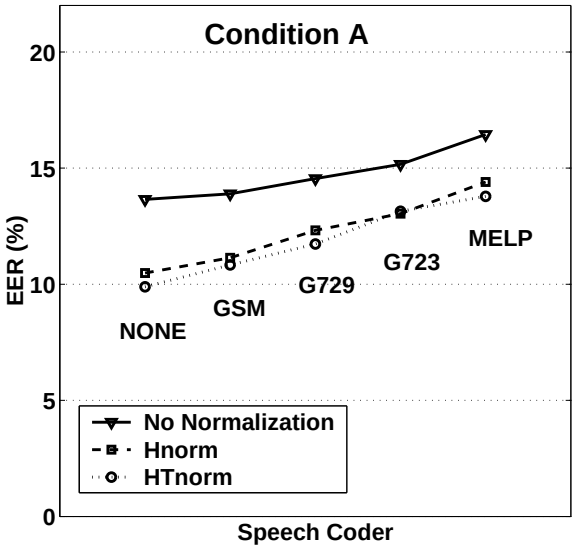

Figure 3: *Equal-Error-Rates for Condition A (matched case where background model, training and testing speech are all coded).*

of removing the handset dependent biases and scales is to normalize the non-target speaker score distributions such that they have zero mean and unit standard deviation for speech from both handset types. This results in better performance when using a single threshold for detection. In addition to removing handset bias and scales, Hnorm also helps normalize log-likelihood scores across different speaker models, again resulting in better performance when using speaker-independent thresholds as in the NIST Speaker Recognition Evaluations [2, 11]. Hnorm is in effect estimating speaker and handset specific thresholds and mapping them into the log-likelihood score domain rather than using them directly.

HTnorm is similar to Hnorm but with the following difference: in Hnorm we compute normalization parameters (means and variances) *for each speaker model* by scoring a fixed set of imposter test segments with that particular speaker model, while in HTnorm the normalization parameters are computed *for each test message* by scoring that particular test message with a fixed set of imposter speaker models.

3. Experiments and Results

The data used for this paper are derived from NIST Speaker Recognition Benchmarks [3] (SRB) which are made up of conversational telephone speech. Coded speech is generated by encoding and then decoding the speech segments with each of the four speech coders, simulating the condition where speech originating from a land line is encoded digitally during transmission. The data for training the UBM and for computing Hnorm and HTnorm parameters was selected from the 1999 NIST SRB. The test data used in this paper are that of the single speaker detection task in the 2000 NIST SRB. This is a much larger and more demanding data set than was used in [4] where the handset was not allowed to vary between training and testing and only electret handsets were used. In contrast, the 2000 NIST SRB includes tests where the training and testing handsets are different and it includes both electret and carbon-button handsets. In addition there are over 6000 test utterances compared with the 625 test utterances used in [4]. The results reported below reflect a pooling of all single-speaker test segments from both male and female speakers. Although both genders are included in the results there are no cross-gender tests as such tests significantly reduce the difficulty of the problem.

The application of Hnorm and HTnorm to coded speech requires that the handset type (CARB or ELEC) be determined from the coded speech. In the experiments below, the handset type for coded speech was detected using GMMs trained from uncoded CARB and ELEC speech with a detection threshold adjusted for the coder as in [12].

3.1. Test Conditions

There are three places in the system where coded speech may be encountered: the test data, the target speaker training data, and the background model training data. In the

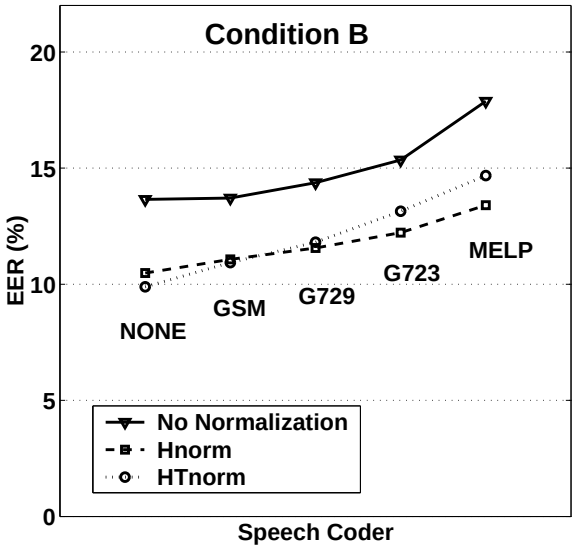

Figure 4: *Equal-Error-Rates for Condition B (partially mismatched case where the background model and training speech are coded and the testing speech is not coded).*

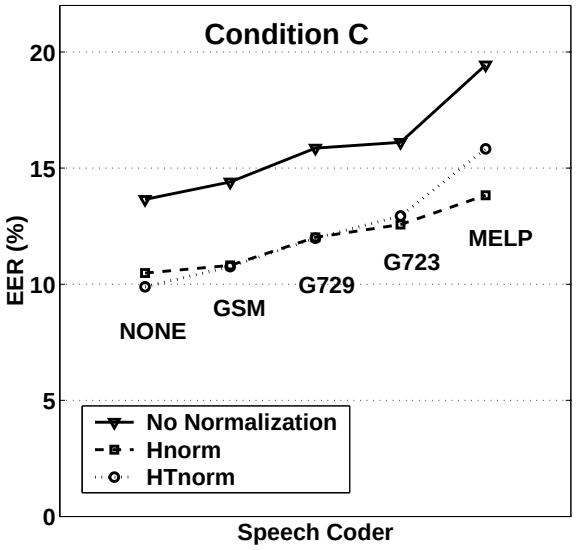

Figure 5: *Equal-Error-Rates for Condition C (partially mismatched case where only the test speech is coded and the background model and training speech are not coded).*

matched case all of the speech is coded while in the mismatched cases some of speech is coded while some of the speech is not coded. For each speech coder, there are four conditions that we tested and compared to a baseline condition in which no coding was performed. We describe the conditions in order of decreasing degree of matching between training and testing.

- **Condition A:** This is the *fully matched* case where background and target models are derived from coded speech and the test data is also coded. In this case training and testing speech are coded and a matching coded UBM is available.
- **Condition B:** This is a *partially mismatched* case where the UBM and target model are derived from coded speech and the test data is from uncoded speech. Since the uncoded test messages are scored against two coded models, we expect performance to decrease relative to condition A. In this case training and testing speech are mismatched but a UBM matching the training speech is available.
- **Condition C:** This is a *partially mismatched* case where the UBM and target model are derived from uncoded speech and the test data is from coded speech. Since the coded test messages are scored against two uncoded models, we expect performance similar to condition B. As in condition B the training and testing speech are mismatched but a UBM matching the training speech is available.
- **Condition D:** This is the *fully mismatched* case where the background model is derived from uncoded speech and the target models and test data are from coded speech. The test data is thus scored against one model derived from coded speech (the target speaker model) and one model derived from uncoded speech (the background model); as such, we expect the worst performance for this case. In this condition the training and testing speech are matched but a matching coded UBM is not available.

In each condition, the imposter test segments used to compute Hnorm parameters were matched to the test speech segments and the imposter models used for HTnorm were trained in the same manner as the target speaker models. For example, in condition D, the test segments are coded and therefore the imposter test segments used to compute Hnorm parameters are also coded. Likewise, in condition D the target models are trained using coded speech and a clean speech UBM so the imposter models for HTnorm are also trained from coded speech and a clean speech UBM.

3.2. Matched Condition with Score Normalization

Speaker detection performance for the various coders in the fully matched condition is shown in Figure 3. This is the

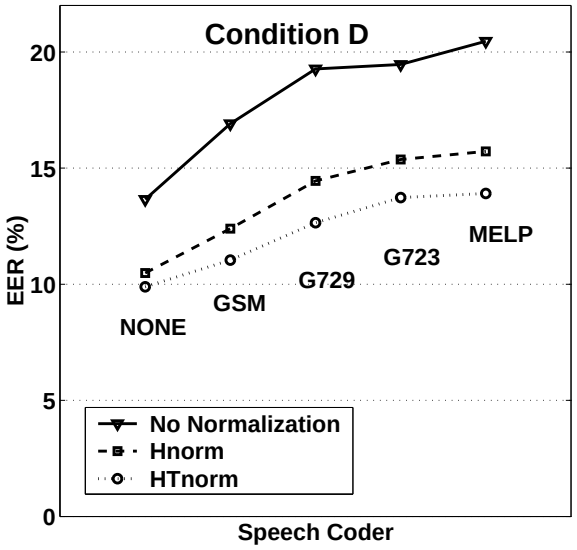

Figure 6: *Equal-Error-Rates for Condition D (fully mismatched case where training and testing speech are coded but the background model speech was not coded).*

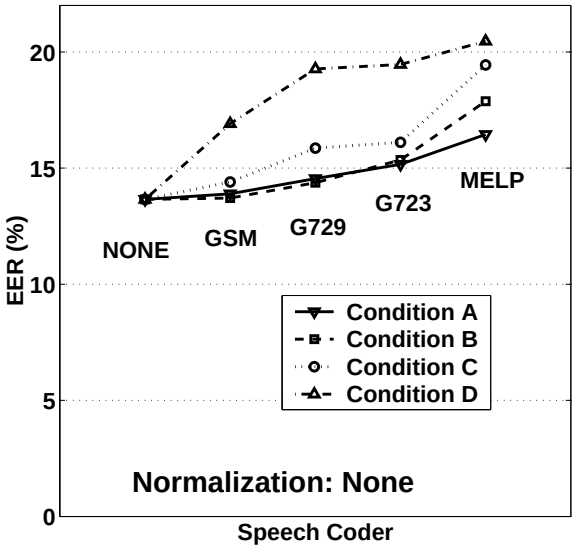

Figure 7: *Equal-Error-Rates for all four conditions (A, B, C, and D) with no score normalization.*

condition where all training and testing speech are coded and a matching coded UBM is used. Performance is shown in terms of the Equal-Error-Rate (EER) which is the operating point where the probabilities of miss and false alarm are equal. Speaker detection performance for the GSM coder is nearly identical to the performance on uncoded telephone speech and there is a slight increase in the EER as the coder bit rate (and speech quality) decreases. The figure shows that both Hnorm and HTnorm are as effective for coded speech as they are for uncoded speech, decreasing the EER about 2-3%. This is true even for G.723 and MELP where the handset identification error is roughly double the error for uncoded speech [12].

3.3. Mismatched Conditions

Speaker detection performance in the various mismatched conditions is shown in Figures 4, 5, and 6. The trends seen here are similar to the matched condition where both Hnorm and HTnorm significantly improve system performance. In the partially mismatched conditions (B and C) HTnorm appears to be as effective as Hnorm is for the high-rate coders GSM and G.729, but for the low-rate MELP coder HTnorm is not as effective as Hnorm. In the fully mismatched condition (D) the use of score normalization (either Hnorm or HTnorm) yields greater performance gains that it does in the matched (A) or partially mismatched (B and C) conditions. It is also notable that in the fully mismatched condition HTnorm has significantly better performance than Hnorm.

All of the conditions (A, B, C, and D) are directly compared without score normalization in Figure 7. The figure shows that in conditions A and B the performance is identical (except for the MELP coder), in condition C there is a slight performance drop, and in condition D there is a large performance drop. When Hnorm is applied, as shown in Figure 8, conditions A, B, and C have similar performance but condition D still shows a performance drop. When HTnorm is used, as shown in Figure 9, there is little to no difference between the four conditions, except for the MELP coder in conditions B and C.

4. Summary

In this paper, we demonstrated that the adapted GMM-UBM speaker recognition system can be effectively used for text-independent speaker detection when telephone speech has been compressed using common speech coding algorithms. There is only a slight increase in the EER for toll quality speech coders (GSM and G.729) as compared with the baseline uncoded speech and the performance loss for lower rate speech coders (G.723 and MELP) is only slightly greater. It was shown that score normalization techniques such as Hnorm and HTnorm can be applied to coded speech and that both score normalization techniques are as effective at improving system performance for coded speech as they are for uncoded speech. Overall, the effect of speech coding on the adapted GMM-UBM speaker recognition system is

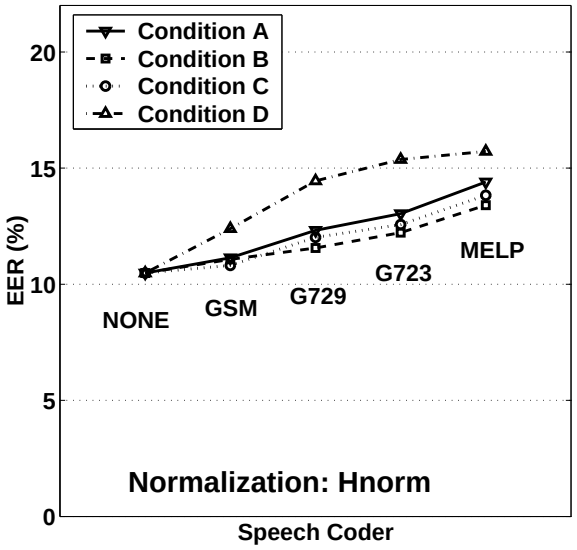

Figure 8: *Equal-Error-Rates for all four conditions (A, B, C, and D) with Hnorm.*

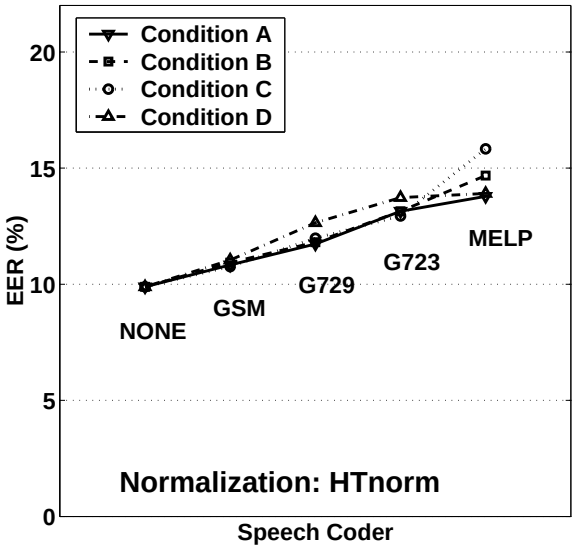

Figure 9: *Equal-Error-Rates for all four conditions (A, B, C, and D) with HTnorm.*

relatively benign under most conditions. Although there is a significant performance loss if the speech used to train the background model was not processed through the same speech coder as the the speech used to train speaker models (condition D), this performance loss can be eliminated if HTnorm is used.

5. References

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing*, Vol. 10, No. 1-3, pp. 19-41, January/April/July 2000.

[2] A Martin, M. Przybocki, "The NIST 1999 Speaker Recognition Evaluation – An Overview", *Digital Signal Processing*, Vol. 10, No. 1-3, pp. 1-18, January/April/July 2000.

[3] Linguistic Data Consortium, http://www.ldc.upenn.edu; 1996-1999 NIST Speaker Recognition Benchmarks.

[4] T.F. Quatieri, E. Singer, R.B. Dunn, D.A. Reynolds, and J.P. Campbell, "Speaker and Language Recognition using Speech Codec Parameters," *Proc. Eurospeech '99*, Vol. 2, pp. 787-790, September 1999.

[5] European Telecommunication Standards Institute, "European digital telecommunications system(Phase2); Full rate speech processing functions (GSM 06.01)," ETSI, 1994.

[6] ITU-T Recommendation G.729, "Coding of speech at 8 kb/s using conjugate-structure algebraic-code-excited linear prediction," June 1995.

[7] ITU-T Recommendation G.723.1, "Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kb/s," March 1996.

[8] A.V. McCree, K.K. Truong, E.B. George, T. Barnwell, and V.R. Viswanathan, "A 2.4 kbit/s MELP Coder Candidate for the New U.S. Federal Standard," *Proc. ICASSP '96*, Vol. 1, pp. 200-204, May 1996.

[9] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," *Proc. Eurospeech '97*, pp. 963-966, September 1997.

[10] D. A. Reynolds, "HTIMIT and LLHDB: Speech corpora for the study of handset transducer effects", *Proc. ICASSP '97*, pp. 1535-1538, April 1997.

[11] National Institute of Standards and Technology, http://www.itl.nist.gov/iaui/894.01/tests/spk/index.htm; NIST Coordinated Speaker Recognition Evaluations

[12] R.B. Dunn, T.F. Quatieri, D.A. Reynolds, and J.P. Campbell, "Speaker Recognition from Coded Speech in Matched and Mismatched Conditions," *Proc. Speaker Recognition Workshop '01*, Crete, Greece, pp. 115-120, June 1999.