On the Use of Quality Measures for Text-Independent Speaker Recognition

D. Garcia-Romero, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez and J. Ortega-Garcia

Speech and Signal Processing Group (ATVS)
Department of Audiovisual Engineering and Communications (DIAC)
Technical University of Madrid (UPM), Spain
dgromero@atvs.diac.upm.es, {jfierrez,jgonzalez,jortega}@diac.upm.es

Abstract

The use of quality information on automatic recognition systems is studied. From an apparent definition of what constitutes a quality measure, a framework for the successful exploitation of the quality information is derived. Potential applications are also introduced at different phases of the recognition process, namely: enrollment, scoring and multi-level fusion stages. Traditional likelihood scoring stage is further developed providing guidelines for the practical application of the proposed ideas. Preliminary experiments corroborate the benefits of the proposed quality-guided recognition approach. In particular, a frame-level quality measure meeting a goodness criterion based on deviation from the fundamental frequency is used, obtaining encouraging initial results.

1. Introduction

One of the key points addressed nowadays by automatic Speaker Recognition research is the exploitation of multi-level information in the speech signal [1, 2, 3]. This idea is founded in self-observation and experience since listeners rely on several types or levels of information in the speech signal to recognize speaker's identity [1]. In the same way, it is also observable that humans are able to perform a number of sophisticated tasks, related to the quality of the information available and the sources of that information, when attempting to make a decision. For example, if a person is to make a decision about the identity of a speaker, based on a noisy and low fidelity speech recording, it is logical to think that the portions of the recording less corrupted by the noise should have a higher influence in the final decision. Furthermore, if the person has to make the decision based on the judgement of two experts, it is highly probable that the person would assign different credibilities to each expert depending on the quality of their opinions.

Based on these intuitive ideas, the aim of this paper is to provide a generic framework in which it is possible to incorporate the described capacities to improve the recognition performance.

Previous work in biometrics has shown promising results when incorporating quality measures into the recognition process [4]. New research efforts are also dedicated to the establishment of objective quality measures of biometric traits such as fingerprint [5] and speech signals [6]. Other applications concerned with quality estimations in the field of speaker recognition include model quality assessment [7] and quality-based feature selection as proposed in [1].

2. Theoretical framework

The concept of quality may be defined[1] as the degree of goodness of an element given a certain criterion. Consequently, a quality measure function $Q^{\xi}(\cdot)$ may be formulated as:

$$Q^{\xi}(Y) = p(Y \text{ meets } \xi) \tag{1}$$

where ξ is a specific goodness criterion for Y. As a result of this formulation, $Q^{\xi}(\cdot)$ is bounded and takes values in the range $[0-1]$. Hence, a reliable quality measure function should be able to quantify the quality of Y with a value of 1 when Y totally satisfies ξ and with a value of 0 when Y does not meet the established goodness criterion at all.

Once a general definition of what constitutes a quality measure is established, a framework results in which it is possible to explore its potential uses throughout the speaker recognition process. The crucial benefits brought into the recognition process by knowing the quality of the elements involved are significant, since this information allows the system to be dynamically adjusted. Examples include the importance given to certain portions of the incoming speech signal during the computation of its likelihood or even how the system relies on each of the scores produced by the different levels of information conveyed in the speech signal.

To some extent there might be some confusion between the well known concept of confidence measure,

This work has been supported by the Spanish Ministry for Science and Technology under projects TIC2003-09068-C02-01 and TIC2003-08382-C05-01. J. F.-A. also thanks Consejeria de Educacion de la Comunidad de Madrid and Fondo Social Europeo for supporting his doctoral research.

[1] Cambridge Klett Dictionary.

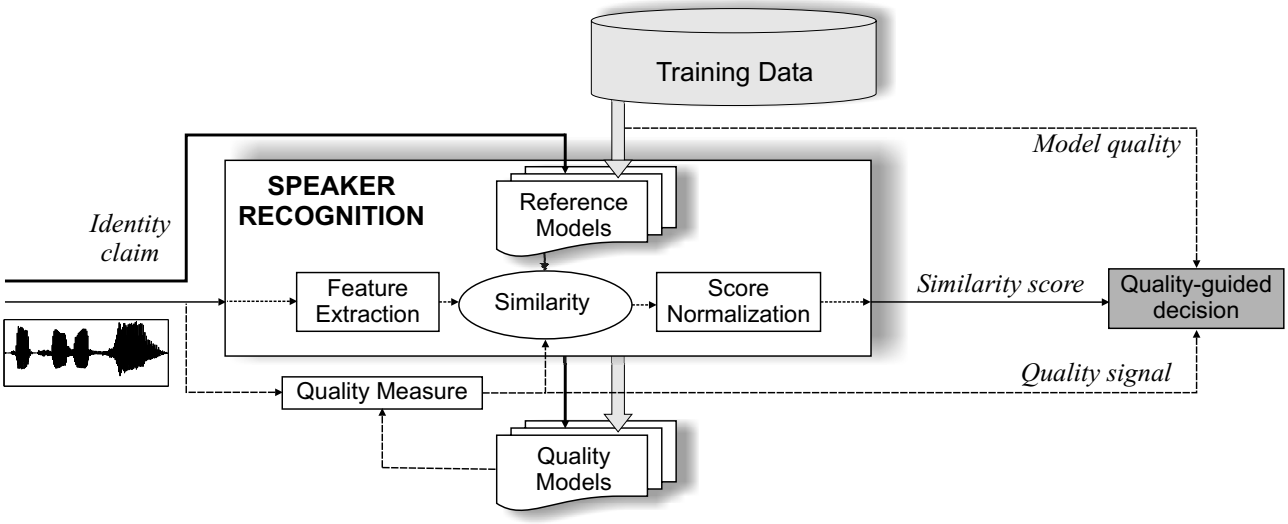

Figure 1: General system model for speaker recognition based on quality measures.

widely used in ASR [8], and the proposed idea of quality measure since both provide information that may be interpreted as how reliable a certain element involved in the recognition process is. It is important to notice that the essence of these two ideas is substantially different. The main purpose of a confidence measure is to quantify how well a model matches some speech data [9], whereas the goal of a quality measure is to quantify how well a certain goodness criterion is satisfied by an element of the system. Thus, the benefit of assigning a confidence estimate to a decoding is succinctly summarized by the phrase: *"knowing what you don't know"* [9], whereas the benefit of estimating the quality of an element of the recognition process is summarized by the phrase:*"knowing the quality of what you have"*. Furthermore, the interpretation of a quality measure as an estimation of the reliability of an element is valid and useful, but it is not the only one. Hence, the concept of quality measure should be regarded as a more general idea than the concept of confidence measure. Therefore, its use should not only be restricted to assessing a confidence estimate of any element of the recognition process, but also to providing information useful for the dynamic adjustment of the system components.

In order to incorporate this general concept of quality measure into the specific framework of speaker recognition systems, we can think of Y as any element of the system (e.g, speech signal, scores, models, thresholds, etc.) and ξ as any factor that affects the behavior of Y and hence the system performance (e.g, SNR, amount of data, course of time, etc.). For example, if we are working with data observed in noisy conditions, Y may be considered as the speech signal and ξ as a criterion based on SNR. Consequently, a quality measure may be stated as follows:

$$Q^{\xi=SNR}(Y) = p(Y > \text{noise}) \tag{2}$$

If we consider the noise normally distributed with mean μ_t and variance σ_t, then the quality of the speech signal, $Y = \{y_t;\ t = 1, \ldots, T\}$, could be segmentally computed by means of the resulting expression [10]:

$$\begin{aligned} q_t^{\xi} &= p(y_t > \text{noise}) \\ &= \int_{-\infty}^{y_t} \frac{1}{\sqrt{2\pi}\,|\sigma_t|} \exp\left(-\frac{(\theta-\mu_t)^2}{2\sigma_t^2}\right) d\theta \end{aligned} \tag{3}$$

The resulting quality signal, $Q^{\xi} = \{q_t^{\xi};\ t = 1, \ldots, T\}$ can be used by the speaker recognition system in several useful ways such as: eliminating the portions of the signal with low quality during the score computation or model training, incorporating the quality information in the score computation function, etc.

3. Potential uses

3.1. Model quality

A major source of performance degradation in Speaker Recognition systems is the mismatch between the acoustic conditions encountered during training and those seen during operation [11]. As a consequence, there is a growing interest in algorithms which adapt models parameters to closely match the incoming speech during the testing phase. Ideally, model adaptation should be based upon the portion of data which increases the quality of the model. Moreover, operational Speaker Recognition systems need an indication of the quality of the newly enrolled models to decide whether to re-enroll or request more enrollment material [7]. Hence, Model Quality

measures provide a mean to refine unsupervised model adaptation procedures and also information to guide the decisions to be taken during the enrollment phase.

3.2. Quality-based score computation

The state of the art in Speaker Recognition systems has been widely dominated during the past decade by the UBM-MAP adapted GMM approach working at the short-time spectral level [12]. Recently, new approaches based on Support Vector Machines (SVM) [13] are achieving similar performance, working at the spectral level, and also providing complementary information useful for the fusion of both approaches, thus increasing the performance in an additive way [14]. Furthermore, higher levels of information conveyed in the speech signal have shown promising discriminative capabilities among speakers and are a major goal of present Speaker Recognition research efforts [1].

A common practice shared among all the above mentioned Speaker Recognition techniques is the use of a pre-processing stage in which two major tasks are accomplished: i) the signal is enhanced according to certain criteria (e.g, channel effects removal, noise reduction, etc.) pursuing an increase in the quality of the signal; ii) hard decisions about the correctness of the basic constituting elements of the data are made (e.g, silence removal, non-speech sound rejection, etc.), preserving those pieces of information that satisfy certain criteria and dismissing the remaining.

This pre-processing approach, combined with a conventional scoring mechanism, has the drawback of regarding all the preserved information as equal in terms of importance or quality once the signal has been pre-processed. Therefore it omits, during the score computation process, the fact that both the information concerning speaker identity and the perturbing artifacts are not distributed uniformly along the pre-processed signal [15].

The underlying idea in the Quality Based Score Computation (QBSC) approach suggests that instead of forcing hard decisions, at an early stage of processing, the score calculation procedure should be adapted to incorporate estimated quality measures (carried on during pre-processing) as weighting factors in the score computation process, see Fig. 1.

Although the QBSC concept is applicable to any of the above mentioned techniques used in Speaker Recognition systems, in the following we are going to particularize for the case of GMM's working at the short-term spectral level, since it is the most widely used paradigm for Speaker Recognition [11].

3.2.1. Quality-based GMM score computation

For a D-dimensional feature vector, $\mathbf{o}$, and a weighted linear combination of M unimodal Gaussian densities, $p(\mathbf{o})$, with the parameters of the density model denoted

$$\lambda = \{w_i, \mu_i, \Sigma_i\} \quad i = 1, \ldots, M \tag{4}$$

the likelihood function is defined as

$$p(\mathbf{o}|\lambda) = \sum\nolimits_{i=1}^{M} w_i p(\mathbf{o}|\mu_i, \Sigma_i) \tag{5}$$

Given a sequence of feature vectors, $O = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T\}$, usually assumed independent, and a quality signal

$$Q^{\xi} = \left\{ q_1^{\xi}, q_2^{\xi}, \ldots, q_T^{\xi} \right\} \tag{6}$$

computed through the speech signal Y with a specific goodness criterion ξ, the likelihood of the model λ incorporating the quality measure as a weighting factor is denoted

$$p(O|Q, \lambda) = \prod\nolimits_{t=1}^{T} p(\mathbf{o}_t|\lambda)^{q_t^{\xi}} \tag{7}$$

The log-likelihood is computed as

$$\log p(O|Q, \lambda) = \sum\nolimits_{t=1}^{T} q_t^{\xi} \log p(\mathbf{o}_t|\lambda) \tag{8}$$

Often, the average log-likelihood is used to normalize out duration effects from the likelihood value. This can be accomplished by dividing Eq. (8) by $\sum_{t=1}^{T} q_t^{\xi}$. Since the assumption of independence between the feature vectors is not precise, this scaling factor can be considered a rough duration compensation [12].

If a quality measure that works in spectro-temporal regions (assigns quality values to each feature vector coefficient) is used instead of one that works in temporal regions (same quality assigned to the entire feature vector), conventional missing data approaches, such as bounded marginalization (BMG) or bounded data imputation (BDI), can be used for the likelihood computation [16].

3.3. Quality-based score fusion

In order to successfully exploit the different levels of information conveyed in the speech signal (e.g., lexical, prosodic, acoustic, etc.) [2, 3] efficient score combination methodologies are necessary [1]. This problem can be formulated as the fusion of different machine experts.

Two sound theoretical frameworks for combining different machine experts in an authentication system are described in [17] and [18], respectively. The former is derived from a risk analysis perspective [19] and the later is based on the statistical pattern recognition theory [20]. Both of them concluded (under some mild conditions which normally hold in practice) that the weighted average is a good way of conciliating different experts and they provided experimental evidence on a multimodal authentication system. Interestingly enough, the approach in [17] was further developed in [4] providing

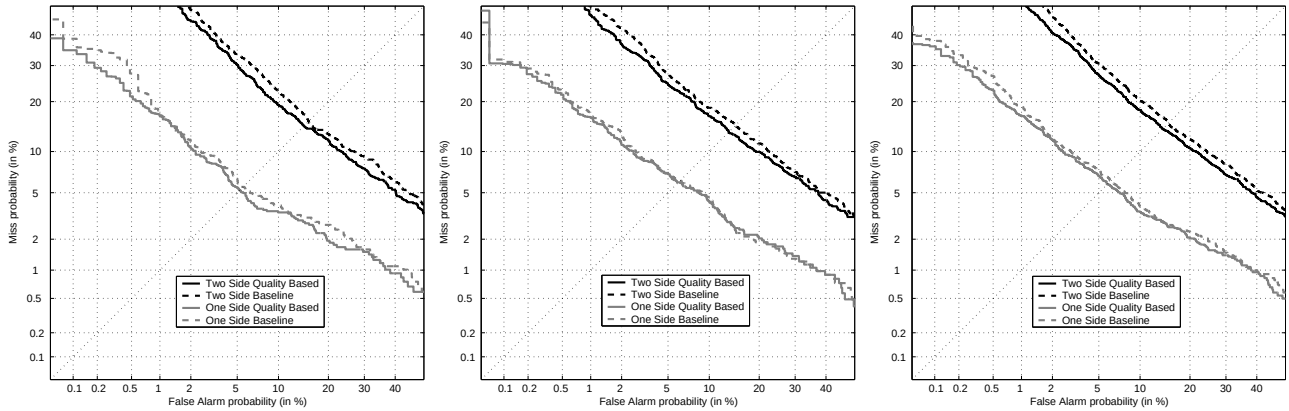

Figure 2: System performance on Switchboard I database for female (left), male (center) and pooling of both (right).

guidelines for the use of quality signals. In particular, a quality-based score fusion scheme is derived in which the weighted average is adapted depending on quality measures of the input biometric samples. Experimental evidence was also reported on a prototype of an authentication application for mobile devices based on voice and fingerprint traits demonstrating the benefits of incorporating quality information at the score fusion level. In the aforementioned work, only quality labels of the fingerprint images were available.

The ideas introduced in this paper can serve as a basis to develop quality labels in case of voice signals. As an example, for a GMM-based speaker verification machine expert, the quality signal in Eq. (6) can be averaged so as to obtain the quality label of the score produced by the voice segment at hand.

4. Experiments

4.1. Databases and experimental protocol

4.1.1. Switchboard I landline database

Partitions 1, 2 and 3 of the Switchboard I database (SWB-I), as defined in [1], have been used for the performance assessment of a speaker verification system on landline telephone data. The number of speaker models involved is 486 (260 male + 226 female). Each target model has been trained with a speech segment of approximately 2.5 minutes comprising one side of a 5 minute telephonic conversation. Two different test sets have been used for the system assessment: i) one side of the conversation test segments (approx. 2.5 min. of speech); ii) two sides of the conversation test segments (approx. 5 min. of speech). The total number of trials obtained with each test set is 8248 (2416 target, 5832 non-target).

4.1.2. NIST 2001 Cellular database

A randomly selected subcorpus of 30 speakers (15 male + 15 female) from the NIST 2001 Cellular database [23] has been established for the preliminary performance assessment of a speaker verification system on cellular data. The total number of trials obtained with this subcorpus is 4215 (281 target, 3943 non-target). Two minutes speech segments were used for target model training and 30 seconds segments for testing. No cross-gender trials have been performed.

4.2. Baseline system

A simplified version (no score normalization has been applied and the number of the GMM mixtures has been reduced to 256) of the ATVS UBM-MAP adapted GMM system [21] has been used to provide a baseline result for comparison with the QBSC adapted implementation of this baseline GMM system.

4.3. Goodness criterion

In order to obtain a quality value for each feature vector, a quality signal was computed at the same frame rate used in the feature extraction process (10ms). A goodness criterion, ξ_{F0} based on $F0$ deviations[2] from the mean, μ_{F0}, was established for that purpose. This criterion is identity-claim dependent (vs. those which are independent of the claimed identity, e.g, SNR) since a model of the $F0$ distribution of the claimed identity is necessary to compute the quality signal. Due to the fact that the $F0$ distribution is Gaussian [22], the training segment of each user was used for the estimation of a user-dependent unimodal gaussian model, $\lambda_{F0} = \{\mu_{F0}, \sigma_{F0}\}$. For each test file, the quality value of each feature vector (belonging to a voiced region of the speech signal) was defined

[2] All $F0$ values are in a logarithmic scale.

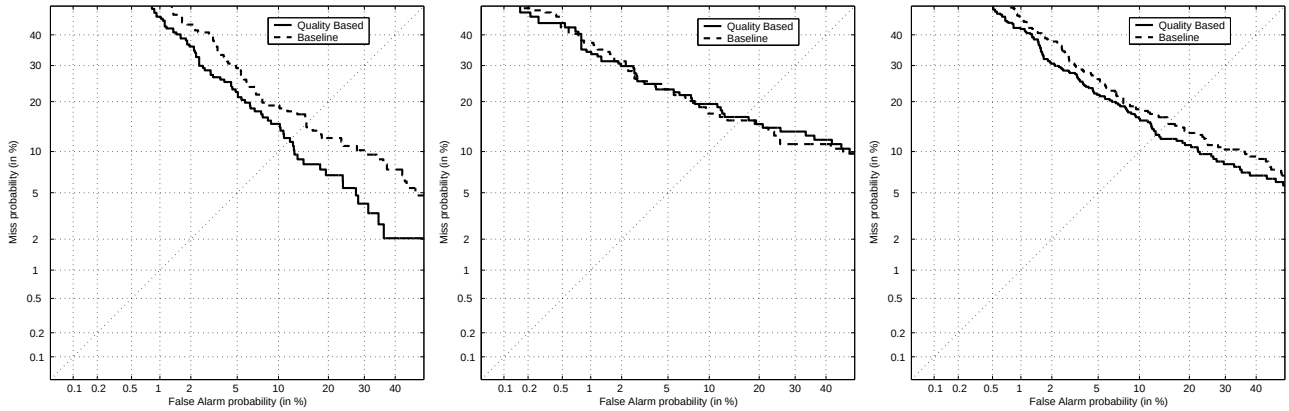

Figure 3: System performance on a subset of NIST 2001 cellular database for female (left), male (center) and pooling of both (right).

at discrete time instant t as

$$q_t^{\xi_{F0}} = p(|y_t^{F0} - \mu_{F0}| < |F0 - \mu_{F0}|) \qquad (9)$$

where $F0 \sim N(\mu_{F0}, \sigma_{F0})$ is the pitch model of the claimed user and y_t^{F0} is the estimated pitch of the test segment at instant t.

For the unvoiced regions of the speech signal a fixed quality value, $q_{t_{unv}}$ is set a priori. In the following experiments this value was fixed to 0.5.

4.4. Results

In Figs. 2 and 3 the performance assessment of both the baseline GMM Speaker Verification system and the QBSC adaptation are depicted in the form of DET plots for the aforementioned corpora.

In relation to the one-side test set of the SWB-I database, see Fig. 2, a slight improvement is obtained by means of using the QBSC adaptation of the baseline system for all the gender partitions. This result is more noticeable in the low false alarm region of the DET curve, and especially in the female partition. Table 1 shows the baseline and QBSC adaptation performance for the EER and DCF[3] operational points.

A bigger improvement is obtained for the two-sides test set, see Fig. 2 and Table 2. The fact that 2 speakers are involved in the two-sides test segments makes this set more suitable for the achievement of better results since a bigger portion of the speech signal may be considered as corrupted. In the special case of a quality measure capable of quantifying the speech segments not belonging to the target speaker with a low quality value, the QBSC adapted system may perform some kind of "speaker spotting". This may be the case for the selected $F0$-based quality criterion since it is possible to discriminate among speakers based on $F0$ information [22]. Therefore, the "speaker spotting" effect of the selected $F0$ quality measure provides a justification for the better performance on the two-sides test set.

[3] As defined in the yearly NIST-SRE [23].

	Baseline		Quality-Based	
Partition	EER(%)	DCF(%)	EER(%)	DCF(%)
Female	5.64	0.0268	5.31	0.0254
Male	6.31	0.0258	6.09	0.0244
Pooled	6.13	0.0277	5.88	0.0257

Table 1: Results on the Switchboard I one-side test set.

	Baseline		Quality-Based	
Partition	EER(%)	DCF(%)	EER(%)	DCF(%)
Female	15.52	0.0682	14.52	0.0636
Male	14.72	0.0593	13.34	0.0548
Pooled	15.09	0.0657	14.01	0.0602

Table 2: Results on the Switchboard I two-sides test set.

Regarding the NIST 2001 Cellular subcorpus, the QBSC adaptation of the baseline system obtains a substantial improvement in the system performance for the female and pooled partitions, remaining almost equal for the male partition. The biggest improvement, see Table 3, is obtained for the female partition, more than 3 points in EER (20% relative improvement). The fact that highly pitch-mismatched speech segments degrade speaker recognition performance, especially for female speakers [24], may be a justification for this result. The QBSC system reduces the contribution of the highly pitch-mismatched regions to the final score value and therefore increases the system performance.

	Baseline		Quality-Based	
Partition	EER(%)	DCF(%)	EER(%)	DCF(%)
Female	15.01	0.0604	11.95	0.0535
Male	15.72	0.0458	16.47	0.0430
Pooled	15.23	0.0537	12.89	0.0486

Table 3: Results on NIST 2001 Cellular subcorpus.

5. Conclusions

An overview of the use of quality information for automatic speaker recognition systems has been reported. From an apparent definition of what constitutes a quality measure, a framework for the successful exploitation of the information conveyed in the estimated quality has been derived. Potential applications have also been introduced at different phases of the recognition process, namely: enrollment, scoring and multi-level fusion stages. Finally, traditional likelihood scoring has been further developed providing guidelines for the practical application of the proposed ideas.

Preliminary experiments on quality-based score computation corroborate the benefits of the proposed quality-guided recognition approach on both landline and cellular data. In particular, a frame-level quality measure meeting a goodness criterion based on deviation from the fundamental frequency has been used. Up to 20% of relative improvement at the EER operational point has been obtained for the female partition of the preliminary NIST 2001 Cellular subcorpus.

6. References

[1] D. A. Reynolds et al., "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition," in *Proc. of ICASSP*, vol. 4, pp. 784–787, 2003.

[2] J. P. Campbell, D. A. Reynolds, and R. B. Dunn, "Fusing high- and low-level features for speaker recognition," in *Proc. of EuroSpeech*, pp. 2665–2668, 2003.

[3] D. Garcia-Romero, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, "Support vector machine fusion of idiolectal and acoustic speaker information in spanish conversational speech," in *Proc. of ICASSP*, vol. 2, pp. 229–232, 2003.

[4] J. Bigun, J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez, "Multimodal biometric authentication using quality signals in mobile communications," in *Proc. of IAPR Intl. Conf. on Image Analysis and Processing, ICIAP*, IEEE CS Press, pp. 2–13, 2003.

[5] D. Simon-Zorita, J. Ortega-Garcia, J. Fierrez-Aguilar, and J. Gonzalez-Rodriguez, "Image quality and position variability assessment in minutiae-based fingerprint verification," *IEE Proc. Vision, Image and Signal Processing, Special Issue on Biometrics on the Internet*, vol. 150, no. 6, pp. 402–408, 2003.

[6] Doh-Suk Kim, "Objective quality assessment of speech," Special Session in *Proc. of ICASSP*, 2004.

[7] Johan Koolwaaij, Lou Boves, Hans Jongebloed, and Els den Os, "On model quality and evaluation in speaker verification," in *Proc. of ICASSP*, pp. 3759–3762, 2000.

[8] M. Siu, H. Gish, and F. Richardson, "Improved estimation, evaluation and applications of confidence measures for speech recognition," in *Proc. of EuroSpeech*, pp. 831–834, 1997.

[9] D. Williams, "Knowing what you don't know: roles for confidence measures in automatic speech recognition," *Ph.D. thesis, University of Sheffield, England*, 1999.

[10] P. Renevey, "Speech recognition in noisy conditions using missing feature approach," *Ph.D. thesis, Ecole Polytechnique Federale de Lausanne*, 2000.

[11] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *Proc. of ICASSP*, vol. 5, pp. 4072–4075, 2002.

[12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.

[13] W. M. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in *Proc. of ICASSP*, vol. 1, pp. 161–164, 2002.

[14] J. P. Campbell, "Advances in speaker recognition: Getting to know you," *Biometric Consortium Conference, MIT Presentation*, 2003.

[15] N. Malayath, H. Hermansky, S. Kajarekar, and B. Yegnanarayana, "Data-driven temporal filters and alternatives to GMM in speaker verification," *Digital Signal Processing*, vol. 10, no. 1, pp. 55–74, 2000.

[16] J. Barker, L. Josifovski, M. Cooke, and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proc. of ICSLP*, vol. 1, pp. 373–376, 2000.

[17] E. S. Bigun, J. Bigun, B. Duc, and S. Fischer, "Expert conciliation for multi modal person authentication systems by Bayesian statistics," in *Proc. of IAPR Intl. Conf. on Audio and Video based Person Authentication, AVBPA*, Springer LNCS, pp. 291–300, 1997.

[18] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 20, no. 3, pp. 226–239, 1998.

[19] E. S. Bigun, "Risk analysis of catastrophes using experts' judgments: An empirical study on risk analysis of major civil aircraft accidents in europe," *European J. Operational Research*, vol. 87, pp. 599–612, 1995.

[20] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, Wiley, 2001.

[21] D. Garcia-Romero, J. Gonzalez-Rodriguez, J. Ortega-Garcia, "ATVS-UPM results and presentation at NIST'2002 speaker recognition evaluation," May 2002.

[22] M. K. Sonmez, L. Heck, M. Weintraub, and E. Shriberg, "A lognormal tied mixture model of pitch for prosody-based speaker recognition," in *Proc. of EuroSpeech*, vol. 3, pp. 1391–1394, 1997.

[23] NIST SRE:, "http://www.nist.gov/speech/tests/spk/"

[24] R.D. Zilca, J. Navratil, and G.N. Ramaswamy, "Depitch and the role of fundamental frequency in speaker recognition," in *Proc. of ICASSP*, vol. 2, pp. 81–84, 2003.