Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 78 (2016) 42–61

www.elsevier.com/locate/specom

Analysis of speech quality measures for the task of estimating the reliability of speaker verification decisions

Jesús Villalba*, Alfonso Ortega, Antonio Miguel, Eduardo Lleida

ViVoLab, Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain

Received 5 June 2015; received in revised form 27 January 2016; accepted 29 January 2016
Available online 5 February 2016

Abstract

Despite the great advances made in the speaker recognition field, like joint factor analysis (JFA) and i-vectors, there are still situations where the quality of the speech signals involved in a speaker verification (SV) trial are not good enough to take reliable decisions. This fact motivated us to investigate speech quality measures that are related to the SV performance. We analyzed measures like signal-to-noise ratio (SNR), modulation index, number of speech frames, jitter, shimmer, or likelihood of the data given the universal background model (UBM), JFA and probabilistic linear discriminant analysis models. Besides, we introduce a novel and promising measure based on the vector Taylor series (VTS) paradigm, used to adapt a clean GMM to noisy speech. We used Bayesian networks to combine these measures and produce a probabilistic reliability measure. We applied it to detect trials badly classified. We trained our Bayesian network on NIST SRE08 distorted with noise and reverberation and evaluated on a distorted version of SRE10. We found that, for noise, the best measures were SNR and modulation index; and for reverberation, the UBM likelihood. VTS based measures performed well for both types of distortions.

Keywords: Speaker verification; Quality measures; Vector Taylor series; Bayesian Network; Reliability; Robustness.

1. Introduction

In recent times, speaker verification (SV) systems have achieved great performance thanks to advanced modeling techniques like joint factor analysis (JFA), Kenny et al. (2008), i-vectors, Dehak et al. (2011), and probabilistic linear discriminant analysis (PLDA), Kenny (2010), which compensate the variability between the recordings of a given speaker. In datasets recorded in, more or less, controlled conditions, like NIST SRE, NIST Speech Group (2010); 2012), or RSR2015, Larcher et al. (2012), these techniques allow to obtain very low error rates (EER $\sim$ 1%). However, we can still find situations where not even these systems can provide reliable decisions. The performance can dramatically drop due to multiple factors: additive noise, reverberation, Ferrer et al. (2011), age, Lei and Hansen (2009), emotional state, Li et al. (2005), language, Lu et al. (2009), short duration of the utterances, Kanagasundaram et al. (2011), etc.

It is well known that additive and convolutional noises greatly affect the distribution of cepstral features and thus, speaker verification, Ferrer et al. (2011). Examples of additive noise are heating, ventilation and air conditioning (HVAC), a car engine, a second speaker next to the target speaker, voices in a crowded place, etc. On the other hand, convolutional noise or reverberation depends on the physical characteristics of the room where the voice is recorded as well as on the frequency response of the transmission channel.

Emotion mismatch between enrollment and test segments can also damage performance. The work in Li et al. (2005) treats this problem using statistical prosodic patterns of emotional utterances to transform the neutral enrollment speech and train a different speaker model for each emotion.

The effect of age is also analyzed on several works. Lei and Hansen (2009) add a term to JFA to account for age variability achieving some improvement on NIST SRE08. Kelly and Harte (2011) measure the effect of age on speaker recognition by using recordings of celebrities in a time span of 30

* Corresponding author. Tel.: +34647155142.
E-mail addresses: villalba@unizar.es (J. Villalba), ortega@unizar.es (A. Ortega), amiguel@unizar.es (A. Miguel), lleida@unizar.es (E. Lleida).

http://dx.doi.org/10.1016/j.specom.2016.01.005
0167-6393/© 2016 Elsevier B.V. All rights reserved.

years. They concluded that the SV score of the target trials starts to degrade when the distance between enrollment and test exceeds 5 years. Besides, they observed that the score drop-off accelerates when the subject is over 60 years.

The effect of language on performance is shown in Villalba et al. (2008), where the EER in NIST SRE08 mixed language trials degraded by 77% w.r.t. English trials. Lu et al. (2009) propose to add language factors to the JFA model to compensate for language variability.

Speaker verification is not only affected by the mismatch between enrollment and test segments but also by the mismatch between development and evaluation data. That means that the speech that we employ to train UBM, JFA, i-vector extractors, PLDA and score calibration needs to be similar to the speech of the enrollment and test recordings. For example,Lu et al. (2012) address the problem of having noisy tests by adding noise to the training of the the i-vector extractor and PLDA. For this reason, measuring the similarity between development and evaluation data we could predict the reliability of the SV scores.

Several works propose methods to decide the reliability of speaker verification decisions. In many cases, they build on studies carried out in the field of speech recognition, Hansen and Arslan (1995). We can divide these methods into three groups. First, we find approaches based on deriving some confidence from the SV score. The SV score is itself a reliability measure. The higher the score, the more reliable the target decision and vice versa. If the score is a well-calibrated likelihood ratio, we can say that scores near zero indicate that the trial is non-reliable, Brummer and Preez (2006). Bengio et al. (2002) propose to compute the difference between the likelihood of the score given the target and non-target distributions, which are trained on a development set. Poh and Bengio (2005) defined another measure as the difference between the miss rate and false acceptance rate for a certain score (taken as threshold). Thus, the closer the score to the EER operating point, the lower the confidence. Furthermore, Mengusoglu (2004) uses the correlation coefficients between the target and non-target score distributions and defines a confidence measure as the difference between both coefficients.

A second group of works base the confidence on auxiliary information computed from the speech utterances. This information is usually referred as quality measures in the literature. Examples of quality measures are utterance duration and signal-to-noise ratio (SNR), given that it is well known that short utterances and noisy environments reduce the speaker recognition accuracy. For example, Garcia-Romero et al. (2006) uses measures like like SNR and the ITU P.563 objective speech quality assessment, ITU-T (2004), and UBM log-likelihood to implement a quality based fusion scheme. Richiardi and Drygajlo (2008) propose to use high-order statistics of speech such us skewness and kurtosis. Moreover, Harriero et al. (2009) analyzed SNR, ITU P.563, UBM log-likelihood and kurtosis of the LPC coefficients. The authors observe a clear correlation between EER and their quality measures on NIST SRE 2006 and 2008.

Finally, the third group combines the quality measures and the classifier output into a unique measure of reliability. For example, Campbell et al. (2005) feed the SV score, numerator and denominator of the likelihood ratio, SNR, utterance duration and channel labels into a multilayer perceptron to obtain a confidence for each score. Richiardi et al. (2005) apply Bayesian networks (BN) to obtain a probabilistic reliability measure. The BN establishes the causal relationships between the random variables involved in the SV process such as the SV score, quality measures, trial label, trial decision and reliability. These relationships facilitate computing the posterior probability for the trial reliability. In Richiardi et al. (2006a); 2006b), the BN based approach is compared with the previous works, above enumerated. The authors conclude that Bayesian networks outperform previous approaches given the possibility of integrating multiple sources of information.

In Villalba et al. (2012), we revisited Richiardi's work focusing on the analysis of the dependencies between the variables of the Bayesian network. In this paper, we extend our previous work. However this time, we focus on comparing a larger number of quality measures, some of them being novel contributions of this work. Again, we use Bayesian networks to infer the reliability. We mainly focus on distortions derived from the recording channel or device, like additive noise and reverberation. The quality measures selected are related with this type of scenarios.

We intend to use the reliability measure to discard unreliable trials, that is, instead of classifying them as target or non-target, we say that the speaker verification decisions are not trustworthy. Companies dedicated to commercialize speaker verification can benefit from this work. It is useful for applications that must provide very accurate decisions but that do not need to provide decisions for all the trials. An example could be a forensic application where we have several recordings that can prove the guilt of a criminal. The verdict of the court should be only based on the ones that provide a reliable evidence. Another application is telephonic access to bank accounts where, in case of determining that the utterance is unreliable, we can ask the client to repeat the sentence.

In this setup, we have two scores, the speaker verification score and the reliability score. We are aware that approaches where both scores are unified into a unique likelihood ratio may seem a more natural way of addressing this problem. However, for certain commercial applications, having two scores is useful. This allows to distinguish whether the trial is rejected because of the quality of the audio signal or because of other reasons.

This paper is organized as follows. Section 2 describes the quality measures that we used in our experiments. They include modulation index, signal-to-noise ratio, number of speech frames, jitter, shimmer, likelihood of the speech frames given the UBM and factor analysis models, and likelihood of the i-vector given the PLDA model. Besides, we present novel features obtained from the parameters needed to adapt a clean GMM to a noisy signal by applying the vector Taylor series paradigm, Li et al. (2009). We also present a method to