Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 76 (2016) 61–81

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Linguistically-constrained formant-based i-vectors for automatic speaker recognition☆

Javier Franco-Pedroso*, Joaquin Gonzalez-Rodriguez

ATVS - Biometric Recognition Group, EPS, Universidad Autonoma de Madrid, c/Francisco Tomas y Valiente 11, Madrid 28049, Spain

Received 29 April 2015; received in revised form 5 November 2015; accepted 13 November 2015
Available online 1 December 2015

Abstract

This paper presents a large-scale study of the discriminative abilities of formant frequencies for automatic speaker recognition. Exploiting both the static and dynamic information in formant frequencies, we present linguistically-constrained formant-based i-vector systems providing well calibrated likelihood ratios per comparison of the occurrences of the same isolated linguistic units in two given utterances. As a first result, the reported analysis on the discriminative and calibration properties of the different linguistic units provide useful insights, for instance, to forensic phonetic practitioners. Furthermore, it is shown that the set of units which are more discriminative for every speaker vary from speaker to speaker. Secondly, linguistically-constrained systems are combined at score-level through average and logistic regression speaker-independent fusion rules exploiting the different speaker-distinguishing information spread among the different linguistic units. Testing on the English-only trials of the core condition of the NIST 2006 SRE (24,000 voice comparisons of 5 minutes telephone conversations from 517 speakers -219 male and 298 female-), we report equal error rates of 9.57 and 12.89% for male and female speakers respectively, using only formant frequencies as speaker discriminative information. Additionally, when the formant-based system is fused with a cepstral i-vector system, we obtain relative improvements of ∼6% in EER (from 6.54 to 6.13%) and ∼15% in minDCF (from 0.0327 to 0.0279), compared to the cepstral system alone.

Keywords: Automatic speaker recognition; Formant frequencies; Formant dynamics; Linguistically-constrained systems.

1. Introduction

Most of the studies in automatic speaker recognition over the last two decades have been based on compact representations of the speech signal in short analysis windows (i.e. MFCC, RASTA-PLP, etc.) (Kinnunen and Li, 2010). Although they are based on spectral representations of the speech signal, it is difficult to directly relate the physiological traits of an individual with the set of such extracted features due to the additional transformations to which they are subjected (inverse FFT, DCT, etc.) (Darch et al., 2005). Moreover, it is hard to interpret such kind of coefficients inasmuch as they do not correspond to any physical magnitude but to mathematical abstractions (the so-called cesptral domain). Formant frequencies, on the other hand, represent the resonant frequencies of the vocal tract of an individual, being easily interpretable and directly related with anatomical and physiological characteristics (Nolan and Grigoras, 2005; Rose, 2002). This makes them specially suitable for forensic purposes (Becker et al., 2008; Gonzalez-Rodriguez et al., 2007), where formant measurements have been used for forensic voice comparison for several decades (Nolan, 1983; Rose, 2002).

Voice comparison is usually performed in the context of linguistic units in forensic-phonetics (McDougall, 2006), but reported studies are usually based on limited experimental frameworks (in terms of number of speakers, number of analysed linguistic-units, or both) due to the manual processes involved in order to extract formant frequencies or labelling the analysed units. So, it is of broad interest to analyse the abilities of formant frequencies for speaker recognition following a similar approach but applied on a large-scale experimental framework with the aid of fully automatic systems. In this way, the presented results can give useful insights for the practitioners in that field.

☆ Non-standard abbreviations: NIST: US National Institute of Standards and Technology. SRE: Speaker Recognition Evaluation. ASR: Automatic Speech Recognition.

* Corresponding author. Tel.: +34660270705.
E-mail addresses: javier.franco@uam.es, javier.franco.pedroso@gmail.com (J. Franco-Pedroso), joaquin.gonzalez@uam.es (J. Gonzalez-Rodriguez).

http://dx.doi.org/10.1016/j.specom.2015.11.002
0167-6393/

Furthermore, interpretable features are helpful in order to correlate with human observations and may lead to find some clues that could be hidden even for very complex cepstral-based systems (Gonzalez-Rodriguez et al., 2014). Such kind of interpretable features, or the systems that make use of them, are usually classified as *higher-level* (Shriberg, 2007), and sometimes involve some kind of *constraints* (Bocklet and Shriberg, 2009) that are applied either in the feature extraction process (in order to define the feature itself), in the speaker modelling process (in order to reduce the intra-speaker variability), or both of them (Shriberg, 2007). *Higher-level* systems provide very useful and complementary information that usually leads to performance improvements when they are combined with short-term acoustic systems (Dehak et al., 2007b; Kockmann et al., 2010; Reynolds et al., 2003).

With the objective of using interpretable features as formant frequencies but being able to evaluate them in the same challenging conditions of the state-of-the-art systems (e.g. the NIST Speaker Recognition Evaluations framework), we present in this paper a speaker verification system based on formant frequencies through the combination of different linguistically-constrained i-vector systems. While previous approaches (Dehak et al., 2007b; Franco-Pedroso et al., 2013; Gonzalez-Rodriguez, 2011; Kockmann and Burget, 2008) extract the speaker distinguishing information from formant frequency dynamics through trajectories coding in the context of some linguistic units (phones, diphones, syllables or pseudo-syllables), in this work we address this issue by means of the classical derivative coefficients (Furui, 1981; Soong and Rosenberg, 1988), also known as *delta* (Δ) features, widely used in speech processing (Benesty et al., 2008) in order to account for the dynamic information in the cepstral domain. This approach has the advantage of not reducing each linguistic segment (e.g. phone, diphone, etc.) to a single observation vector, relaxing the previous requirements of training data derived from extracting one single feature vector per linguistic segment.

The rest of the paper is organized as follows. Section 2 presents a brief overview of how formant frequencies have been used for speaker recognition, while Section 3 describes the automatic feature extraction process followed in the proposed approach. Section 4 details how linguistically-constrained i-vector systems are built from formant features with the aid of automatically-generated phonetic labels. Section 5 describes the constraint-selection rules and fusion techniques used in order to combine the linguistically-constrained systems for text-independent speaker recognition. The experimental framework and evaluation metrics are presented in Section 6, including a description of our reference cepstral-based speaker recognition system. Results are shown in Section 7 for both independent linguistically-constrained systems and for several constraint combinations, as well as for the combination of formant and cepstral-based systems. Finally, conclusions are drawn in Section 8 and extended results are reported in a final appendix.

2. Formant frequencies for speaker recognition

Formant frequencies have strong individualization potential (Nolan, 1983) and have been used for forensic voice comparison for several decades (Rose, 2002). Usually, formant centre frequencies are extracted at the temporal midpoint of vowels (Rose and Winter, 2010) reflecting in part certain anatomical dimensions of a speaker as the length and configuration of the vocal tract. Also, the mean frequencies over the time-course of the vowel (Zhang et al., 2008) have been used.

In order to obtain richer representations, frame-by-frame formant-frequency distributions have been modelled through either long-term formant distributions (LTFs) (Nolan and Grigoras, 2005) or multivariate Gaussian mixture models (GMMs) (Becker et al., 2008). It is also common to incorporate formant bandwidth measurements in order to complement the information provided by instantaneous formant frequency values (Becker et al., 2008; Gonzalez-Rodriguez, 2011), as they are also related to vocal tract conditions.

Formant dynamics were also proposed for speaker recognition (McDougall, 2006) under the assumption of presenting higher inter-speaker variability within linguistic units than the static measurements of formant frequencies: while speakers seems to show very similar acoustic properties at moments at which 'phonetic targets' (McDougall, 2006) are achieved (e. g. formant frequencies at a segment's temporal midpoint), much larger differences are exhibited in the ways they move between consecutive targets (Nolan, 2002).

This transitional information is omitted by statistical distributions obtained from frame-by-frame formant frequencies. In order to capture this dynamic information, two main approaches have been used: polynomial fitting (Dehak et al., 2007b; McDougall, 2006) and Discrete Cosine Transform (DCT) (Franco-Pedroso et al., 2013; Morrison, 2009) of formant trajectories over linguistic units. Both approaches compute a fixed number of polynomial or DCT coefficients per trajectory and concatenates the coefficients from the different formant trajectories, yielding a single feature vector that captures the dynamic information of the different formants in a given linguistic unit. In order to define the speech region where formant trajectories are computed, both manual segmentations (mainly in the forensic field) (McDougall, 2006; Morrison, 2009) and automatic speech recognition (ASR) systems (Dehak et al., 2007b; Franco-Pedroso et al., 2013) have been used. Using coded trajectories as feature vectors, speakers have been modelled through multivariate kernel distributions (MVK) (Gonzalez-Rodriguez, 2011; Morrison, 2009) or GMM's (Franco-Pedroso et al., 2013) in a linguistic unit-dependent manner, or by means of joint factor analysis (JFA), compensating for intersession variability, by pooling together trajectories from different units (Dehak et al., 2007b).

Similarly, the approach proposed in this paper is based on formant frequencies, but extracts the dynamic information through derivative coefficients (Furui, 1981; Soong and Rosenberg, 1988) regardless of the linguistic content. These coefficients are also extracted at a frame-by-frame rate and combined with the static information of instantaneous formant frequency values. Then, linguistic units are used as constraints applied to feature vectors in order to develop separate i-vector systems for each linguistic unit, allowing to independently analyse their speaker-distinguishing abilities.