DYNA
ISSN: 0012-7353
ISSN: 2346-2183
Universidad Nacional de Colombia

Porras-Plata, Dagoberto; Sarria-Paja, Milton; Sepúlveda-Sepúlveda, Alexander
Speaker verification system based on articulatory information from ultrasound recordings
DYNA, vol. 87, no. 213, 2020, April-June, pp. 9-16
Universidad Nacional de Colombia

DOI: https://doi.org/10.15446/dyna.v87n213.81772

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE MEDELLÍN
FACULTAD DE MINAS

DYNA
http://dyna.medellin.unal.edu.co/

Speaker verification system based on articulatory information from ultrasound recordings

Dagoberto Porras-Plata [a], Milton Sarria-Paja [b] & Alexander Sepúlveda-Sepúlveda [a]

[a] *Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones (E3T), Universidad Industrial de Santander, Bucaramanga, Colombia. dagoberto.porras@correo.uis.edu.co, alexander.sepulveda@gmail.com*
[b] *Facultad de Ingenierías, Universidad Santiago de Cali, Cali, Colombia. milton.sarria00@usc.edu.co*

Received: August 19th, 2019. Received in revised form: February 11th, 2020. Accepted: February 26th, 2020

Abstract
Current state-of-the-art speaker verification (SV) systems are known to be strongly affected by unexpected variability presented during testing, such as environmental noise or changes in vocal effort. In this work, we analyze and evaluate articulatory information of the tongue's movement as a means to improve the performance of speaker verification systems. We use a Spanish database, where besides the speech signals, we also include articulatory information that was acquired with an ultrasound system. Two groups of features are proposed to represent the articulatory information, and the obtained performance is compared to an SV system trained only with acoustic information. Our results show that the proposed features contain highly discriminative information, and they are related to speaker identity; furthermore, these features can be used to complement and improve existing systems by combining such information with cepstral coefficients at the feature level.

Keywords: speech processing; speaker verification; articulatory parameters; ultrasound; i-vectors; GMMs.

Sistema de verificación de hablantes utilizando información articulatoria de grabaciones de ultrasonido

Resumen
Los sistemas actuales de verificación de hablantes (VH) pueden verse afectados por variaciones inesperadas durante la fase de validación, tales como ruido de entorno o cambios en el esfuerzo vocal. En este trabajo se evalúa la información articulatoria del movimiento de la lengua como medio para mejorar el desempeño de los sistemas de verificación del hablante. Se utilizó una base de datos en español, donde además de las señales de voz, también se adquiere información articulatoria con un sistema de ultrasonido. Se proponen dos grupos de características para representar la información articulatoria y el desempeño obtenido es comparado con un SVH entrenado únicamente con información acústica. Los resultados muestran que las características propuestas contienen gran cantidad de información discriminativa y altamente asociada a la identidad de los hablantes, además que se pueden emplear para complementar y mejorar SVH existentes como por ejemplo combinando dicha información con coeficientes cepstrales.

Palabras clave: procesamiento de señales del habla; verificación de hablantes; parámetros articulatorios; ultrasonido; i-vectors; GMMs.

1. Introduction

Human speech is a natural mode of communication that enables the exchange of highly complex, detailed, or sophisticated information between two or more speakers. This mode of communication not only conveys information related to a message, but also information related to speaker's gender, emotions, or even speaker identity [1]. For many years there has been great interest from scientific community in speech signals, not only from the perceptual point of view but also from acoustics given the number of applications [1-4]. One of such applications is automatic speaker recognition, in which we seek to design a digital system to play the role of a human listener at identifying people based on that

How to cite: Porras-Plata, D, Sarria-Paja, M. and Sepúlveda-Sepúlveda, A, Speaker verification system based on articulatory information from ultrasound recordings. DYNA, 87(213), pp. 9-16, April - June, 2020.

person's voice [2,5,6]. Applications with voice biometrics are burgeoning as a secure method of authentication, which eliminates the common use of personal identification numbers, passwords, and security questions.

In automatic speaker recognition, there are two classical tasks that can be performed: speaker identification (SI) and speaker verification (SV). Identification is the task of deciding, given a speech sample, who among a set of speakers said the sample. This is an N-Class problem (given N speakers), and the performance measure is usually the classification rate or accuracy. Verification, in turn, is the task of deciding, given a speech sample, whether the specified speaker really said the sample or not. The SV problem is a two-class problem of deciding if it is the same speaker or an impostor requesting verification. Commonly, SV exhibits greater practical applications related to SI, especially in access control and identity management applications [2,5,7]. In this study, we address the speaker verification problem by using articulatory information.

The general structure of a speaker recognition system can be explained in three functional blocks as follows: *i*) an input block to measure important features parameters from speech signals; *ii*) a feature selection block or dimensionality reduction, which is a stage that relies on a suitable change (simplification or enrichment) of a representation to enhance the class or cluster descriptions [8]; *iii*) and the generalization/inference stage. In this final stage, a classifier/identifier is trained. The training process involves the parameter tuning of models to describe training samples (i.e., features extracted from speech recordings) [5,8,9].

With automatic speaker recognition, short time-based features such as mel-frequency cepstral coefficients combined with Gaussian mixture models (GMM) or latent variable inspired approaches [10,11] have been, for many years, the typical techniques for text-independent speaker recognition. While different approaches have been proposed to add robustness to speaker recognition systems against environmental noise [12], environmental noise and channel variability are still challenging problems. Hence, it is important to explore other information sources to extract complementary information and to add robustness to the systems. In this regard, articulatory information is known to be less affected by noise and contains speaker identity information, so it is prone to be used in automatic speaker recognition tasks [13]. Furthermore, several studies have shown articulatory related information is an important source of variability between speakers [13-15]. Nevertheless, measuring articulatory information is a complex task, and there are alternative strategies that have been proposed to estimate articulatory movements using acoustic information [16-18]. As an example, in [19,20] authors propose to use the trajectories of articulators estimated from the acoustic signal as a method to improve performance in speaker verification systems.

By contrast, there are research works that use direct measurements from different sensor types. For instance, electromagnetic articulography (EMA) data have been used for speaker recognition purposes, but EMA data acquisition is costly. In [21], recordings from 31 speakers were used, 18 males and 13 females, achieving 98.75% accuracy for speaker identification tasks by using EMA data. EMA uses sensor coils placed on the tongue (typically in 2 or 3 points) and other parts of the mouth to measure their position and movement over time during speech, which allows to capture data with great time resolution but limited spatial resolution [22]. Furthermore, articulatory movements can also be measured by using real-time vocal tract magnetic resonance imaging (rtMRI), and results suggest that articulatory data from rtMRI can also be used for speaker recognition purposes [23,24].

As an alternative to the previously described acquisition methods, ultrasound equipment is considerably cheaper and more portable. Furthermore, it can capture in real-time the mid-sagittal plane of the tongue but in a noninvasive way, in contrast to EMA. In addition, ultrasound does not add interference while acquiring acoustic signals as in MRI, and neither does it affect natural movements of articulators, as in EMA and MRI. Hence, ultrasound is a technique of higher cost-benefit when compared with other data acquisition methods for measuring articulatory information [25]. In fact, phonetic research has employed 2D ultrasound for a number of years for investigating tongue movements during speech. Motivated by these advantages, and under the hypothesis that articulatory information can be useful for speaker recognition tasks [20], this work proposes to analyze the potential use of articulatory information associated with the tongue superior contour for speaker verification.

2. Materials and methods

2.1. Data acquisition

Our dataset contains recordings of 17 speakers (9 female and 8 male). The ages of participants range from 20 to 24 years, and participants are from the same geographical location, Bucaramanga in Colombia. All recordings contain information from the acoustic signal and the ultrasound video. For data acquisition we used the ultrasound probe *PI 7.5 MHz Speech Language Pathology,* (https://seemore.ca/portable-ultrasound-products/pi-7-5-mhz-speech-language-pathology-99-5544-can/) manufactured by Interson. To avoid artifacts, we used a helmet *Articulate Instruments* [26] to stabilize the probe [26,27], acoustic signals were recorded using the Shure SM58 microphone at 48000 Hz sampling rate, and the recordings were saved in WAV format. Fig. 1 shows the hardware used in present work. The ultrasound video was stored as a grey scale images sequence using the format JPG, with dimensions of 800x600 pixels. Each speaker uttered 30 Spanish sentences in a total of 340 seconds, and approximately 4000 images per speaker were recorded. For additional details, the interested reader can find information in [28].

2.2. Articulatory features

The articulatory features were computed in two stages

Figure 1. Equipment used for the parallel acoustic-articulatory data collection: microphone, personal computer, ultrasound probe and stabilization headset.
Source: [28].

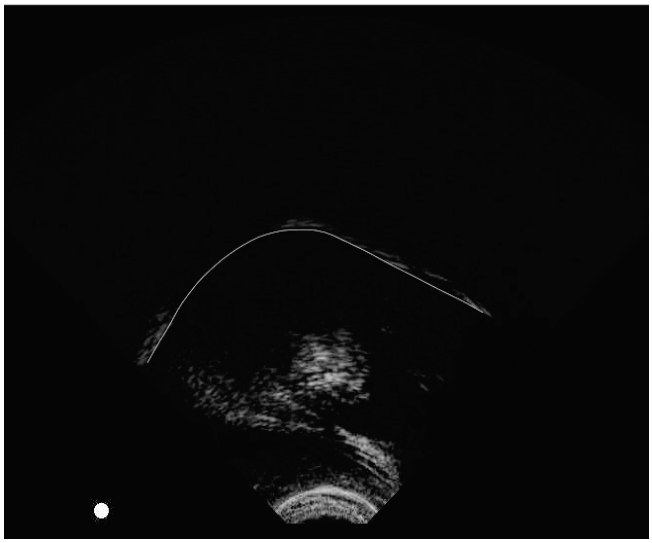

Figure 2. Example: Tongue contour from a 2D ultrasound image.
Source: The Authors.

using the image sequence from the ultrasound video. First, the tongue contour is segmented using the software *EdgeTrak*

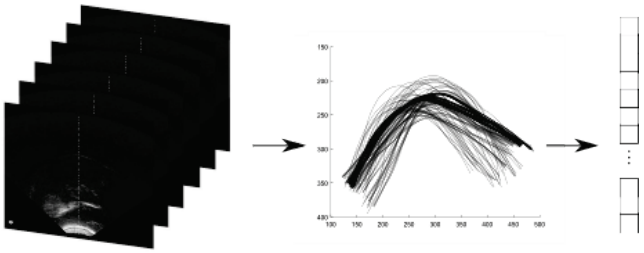

Figure 3. Diagram representing articulatory information (ultrasound images).
Source: The Authors.

[29], which uses an algorithm based on active contours [30]. Fig. 2 shows an example of the extracted contour from an ultrasound image. In several instances, it was necessary to apply a correction over the estimated contour.

Next, having $N = 50$ coordinates (x, y) describing the tongue contour for each frame, they are modeled. Having the tongue contour from an image, we propose to use two approaches:

- Adjust directly a six-degree polynomial function using all $N = 50$ points, as shown in Fig. 3, and append to the 7 coefficients the horizontal coordinates from the initial x_i and final x_f points, resulting in a nine-dimensional feature vector $a_k = [p_0, p_1, \dots, p_6, x_i, x_f]^T$ describing the k-th ultrasound image, where p_j denotes the j-th polynomial coefficient.
- Take five equally spaced points with coordinates $\{x_i, y_i : i = 1{,}10{,}20{,}30{,}40{,}50\}$ and arrange them into a ten dimensional feature vector $b_k = [x_1, x_2, \dots, x_5, y_1, y_2, \dots, y_5]^T$.

2.3. Acoustic features

The most popular analysis method for automatic speech recognition combines cepstral analysis theory [31] with aspects related to the human auditory system [2]. The so-called mel-frequency cepstral coefficients (MFCC) are the classical frame-based feature extraction method widely used in speech applications. Originally proposed for speech recognition, MFCC also became a standard for many speech-enabled applications, including speaker recognition [2, 32, 33]. In addition to the spectral information represented by the MFCCs, the temporal changes in adjacent frames play a significant role in human perception [34].

For MFCC computation, each speech recording is pre-emphasized and windowed in overlapped frames of length τ using a Hamming window to smooth the discontinuities at the edges of the segmented speech frame. Let *x(n)* represent a frame of speech that is pre-emphasized and Hamming-windowed. First, *x(n)* is converted to the frequency domain by an *N* point discrete Fourier transform (DFT). Next, *P* triangular bandpass filters spaced according to the mel scale are imposed on the spectrum. These filters do not filter time domain signals, they instead apply a weighted sum across the *N* frequency values, which allows to group the energy of

frequency bands into P energy values. Finally, a discrete cosine transform (DCT) is applied to the logarithm filter bank energies resulting in the final MFCC coefficients. This process is repeated across all the speech frames, hence each speech recording is represented as a matrix of MFCC coefficients, being each arrow the respective coefficients of a single frame.

The temporal changes in adjacent frames play a significant role in human perception [34]. To capture this dynamic information in the speech, first- and second-order difference features (Δ and ΔΔ MFCC) can be appended to the static MFCC feature vector. In our experiments, 27 triangular bandpass filters spaced according to the mel scale were used in the computation of 13 MFCC features, and Δ and ΔΔ features were appended, resulting in a 39-dimensional feature vector. The interested reader is referred to [2, 33, 35] for more complete details regarding the computation and theoretical foundations of the MFCC feature extraction algorithm. The MFCC were computed on a per-window basis using a 25 ms hamming window with 10 ms overlap. We used the HTK toolkit to compute these features [36]. After dropping frames where no vocal activity was detected, cepstral mean and variance normalization was applied per recording to remove linear channel effects [37,38].

2.4. GMM Speaker modelling

For speaker modeling, the well-known Gaussian mixture model has remained in the scope of speaker recognition research for many years. The use of Gaussian mixture models for speaker recognition is motivated by these models' capability to model arbitrary densities, and the individual components of a model are interpreted as broad acoustic classes [5]. A GMM is composed of a finite mixture of multivariate Gaussian components and the set of parameters denoted by λ. The model is characterized by a weighted linear combination of C unimodal Gaussian densities by the function:

$$\mathrm{p}(\mathrm{o}|\lambda) = \sum_{\mathrm{i}=1}^{\mathrm{C}} \alpha_{\mathrm{i}} \mathcal{N}(\mathrm{o}, \mu_{\mathrm{i}}, \Sigma_{\mathrm{i}}) \quad (1)$$

where o is a D-dimensional observation or feature vector, α_i is the mixing weight (prior probability) of the i-th Gaussian component, and $\mathcal{N}(\cdot)$ is the D-variate Gaussian density function with mean vector μ_i and covariance matrix Σ_i.

Let $\mathcal{O} = \{o_1, \ldots, o_k\}$ be a training sample with K observations. Training a GMM consists of estimating the parameters $\lambda = \{\alpha_i, \mu_i, \Sigma_i\}_{i=1}^{C}$ to fit the training sample $\mathcal{O}$ while optimizing a cost function. The typical approach is to optimize the average log-likelihood (LL) of $\mathcal{O}$ with respect to the model λ and is defined as [8]:

$$LL(\mathcal{O}, \lambda) = log p(\mathcal{O}|\lambda) = \frac{1}{K} \sum_{k=1}^{K} log \sum_{i=1}^{C} \alpha_i \mathcal{N}(o_k, \mu_i, \Sigma_i) \quad (2)$$

The higher the value of LL, the higher the indication that the training sample observations originate from the model λ. Although gradient-based techniques are feasible, the popular expectation-maximization (EM) algorithm is used for maximizing the likelihood with respect to given data. The interested reader is referred to [8] for more complete details. For speaker recognition applications, first, a speaker-independent *world model* or *universal background model* (UBM) is trained using several speech recordings gathered from a representative number of speakers. Regarding the training data for the UBM, selected speech recordings should reflect the expected alternative speech to be encountered during recognition. This applies to both the type and the quality of speech, as well as the composition of speakers. Next, the speaker models are derived by updating the parameters in the UBM using a form of Bayesian adaptation [39]. In this way, the model parameters are not estimated from scratch, with prior knowledge from the training data being used instead. It is possible to adapt all the parameters, or only some of them from the background model. For instance, adapting only the means has been found to work well for speaker recognition, as was shown in [39], according to the authors, there was no benefit in updating parameters such as covariance matrix or priors associated to each gaussian.

2.5. i-vector modeling approach

The i-vectors extraction technique was proposed in [10] to map a variable length frame-based representation of an input speech recording to a small-dimensional feature vector while retaining the most relevant speaker information. First, a C-Component GMM is trained as an universal background model (UBM) using the Expectation -- Maximization (EM) algorithm and the data available from all speakers from the train set or background data, as described in the previous section. Speaker and session-dependent supervectors of concatenated GMM means are modeled as:

$$M = m + T\phi \quad (3)$$

where m is the speaker- and channel-independent supervector; $T \in \mathbb{R}^{CF \times D}$ is a rectangular matrix of low rank covering the important variability (total variability matrix) in the supervector space; C,F and D represent, respectively, the number of Gaussians in the UBM, the dimension of the acoustic feature vector and the dimension of the total variability space; and finally, $\phi \in \mathbb{R}^{D \times 1}$ is a random vector with density $\mathcal{N}(0, I)$ and referred to as the identity vector or *i-vector* [12]. A typical i-vector extractor can be expressed as a function of the zero- and first-order statistics generated using the GMM-UBM model, and it estimates the Maximum a Posteriori (MAP) point estimate of the variable ϕ. This procedure is complemented with some post-processing techniques such as linear discriminant analysis (LDA), whitening, and length normalization [40].

Table 1.
Number of speakers and recordings in each set.

	UBM	Experiments	
		Enroll	test
# of speakers	15	2	2
# of sessions	All	24	6

Source: The Authors.

3. Experimental setup

During testing, in a verification scenario, we consider a testing sample $\mathcal{O}_t$ and a hypothesized speaker with model λ_{hyp}, and the task of the speaker verification system is to determine if $\mathcal{O}_t$ matches the speaker model, when using a GMM+MAP adaptation based system. When using the *i-vector/PLDA* approach, first it is necessary to extract the i-vector ϕ_{test} from the respective testing sample, $\mathcal{O}_t$. Next, we consider ϕ_{test} and a hypothesized speaker with a set of averaged i-vectors ϕ_{enrol}, which is the speaker model, and the task of the speaker verification system is to determine if ϕ_{test} matches the speaker model. There are two possible hypotheses: 1) The testing sample is from the hypothesized speaker, and 2) the testing sample is not from the hypothesized speaker.

For the experiments herein, we use the probabilistic linear discriminant analysis (PLDA) approach for scoring, as recommended in [41]. The PLDA model splits the total data variability into within-individual (*U*) and between-individual (*V*) variabilities, both residing in small-dimensional subspaces. Originally introduced for face recognition, PLDA has become a standard in speaker recognition. The interested reader is referred to [42] for further details regarding PLDA formulation.

The decision can be made by computing the log-likelihood (score) between the two hypotheses; details can be found in [5,41,42]. For the experiments herein, the dimensionality of LDA transformation matrix *V* and *U* were fixed to the number of train speaker set and, $M = 0$, respectively; where the LDA model is tuned accordingly per feature set.

In this work, we used all sessions from 15 speakers to train the UBM, as well as 24 recordings from two speakers for enrollment, and for testing, we used six recordings per client, see Table 1. Finally, the performance of the system was measured using the EER (Equal Error Rate), a common performance measure for speaker verification systems [7].

Sixteen Gaussian distributions form the GMM of the UBM model, whose parameters were adjusted by the EM (Expectation-Maximization) optimization algorithm. In order to evaluate the method, it is performed cross-validation in respect to speakers (15 for training, 2 for testing). For each cross-validation iteration, 30 phrases belonging to those 15 speakers are included for training the UBM model. In total, 105 experiments were performed, with a single UBM per experiment. m05, m06 speakers were excluded from the training set because they contain less than 30 sentences. For each experiment, 24 sentences belonging to the two remaining speakers were utilized in order to find the

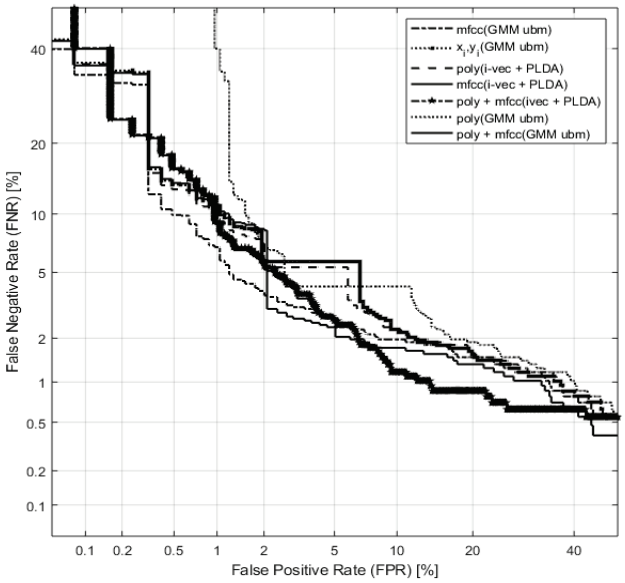

Figure 4. DET curve comparing different speaker verification systems. Source: The Authors.

corresponding two individual models. The remaining 6 sentences per speaker are used to carry out the verification test. It is important to note that for each experiment, we have 10 different ways of selecting the sentences (24 for the individual models and 6 for testing). The final EER value can be estimated by averaging those 1050 EER values corresponding to the tests.

Herein, we implemented two approaches for speaker verification: the classical GMM-UBM+MAP adaptation approach and the most recent *i-vector + PLDA* paradigm; for this purpose, we used the MATLAB tool *MSR Identity Toolbox* [38]. Having acoustic feature vectors such as the MFCC, articulatory features from the polynomial modeling of the tongue contour, and the fusion of these features, we train a GMM as UBM for two purposes: *i)* to apply MAP adaptation and obtain a speaker model, and *ii)* to compute the zero- and first-order statistics (Baum welch statistics) and to train the i-vector extractor. The total variability matrix dimension is set to $D = 16$, and the i-vectors dimension was reduced to 14 using LDA.

4. Results

Table 2 presents the results using articulatory features, and these results are compared with the classical MFCC feature set. In the table, the feature set S1 refers to the polynomial approach, where the feature vector is composed of seven coefficients plus the initial and final horizontal coordinates of the tongue contour $(p_0, p_1, \ldots, p_6, x_i, x_f)$. The feature set S2 refers to the feature set, where we take five equally spaced points and arrange both horizontal and vertical coordinates in a single feature vector $\{x_i, y_i : i = 1{,}10{,}20{,}30{,}40{,}50\}$.

Table 2.
EER (%) comparison of different feature sets using GMM+MAP adaptation and i-vectors + PLDA based systems. Source: Authors

Features	dim	GMM	i-vector + PLDA
S1: $p_0, p_1, \dots, p_6, x_i, x_f$	9	3.73	4.99
S2: $x_1, x_2, \dots, x_5, y_1, y_2, \dots, y_5$	10	5.32	11.51
MFCC	39	2.53	3.69

Source: The Authors.

Table 3.
EER (%) results comparing fusion vs MFCC alone using two different speaker verification systems **S1**: $p_0, p_1, \dots, p_6, x_i, x_f$, **S2**: $x_1, x_2, \dots, x_5, y_1, y_2, \dots, y_5$.

Features	GMM	i-vector + PLDA
MFCCs + S1	3.86	2.11
MFCCs + S2	13.42	15.38
MFCCs	2.53	3.69

Source: The Authors.

As can be seen, best results were obtained using the classical GMM/MFCC paradigm, which achieves an EER = 2.53%. Results achieved by articulatory features, on the other hand, are not as good as results presented by MFCC feature vectors, but these results show that articulatory features contain highly discriminative information associated to speaker identity. As can be seen, the EER = 3.73 achieved by the set S1 has and absolute difference of 1.2% relative to results achieved by MFCC features. It is important to clarify that in the case that S1 and S2 did not contain discriminative information, the verification results would not be comparable with those achieved with the MFCC, which have already been shown as discriminative characteristics in numerous works. These results are in line with previously reported research in [43], where the EER is around 4.7 – 6.8 % using information from the tip and middle of the tongue using an EMA device. It is also important to point out that the GMM+MAP adaptation based systems achieved better performance than the i-vector+PLDA approach. This can happen due to the lack of data for training the i-vector extractor [20], and there is not enough variability to benefit from this technique.

Features described in Sections 2.2 and 2.3 place emphasis on different aspects of the speech production process, thus likely contain complementary information. This hypothesis has motivated the exploration of fusion at different levels to combine the strengths of feature representations extracting complementary information [7]. One way to combine the strengths of these features is by combining them at the frame level by concatenating articulatory features to the MFCC feature vector. For these experiments, it is necessary to synchronize the feature sets to guarantee the temporal information to be aligned. This is done by modifying the frame size for computation of the MFCC to 78 ms.

Results from this fusion strategy are presented in Table 3. As can be seen, important improvements were attained when comparing with results presented in Table 2. In particular, there are some losses with respect to the GMM+MAP adaptation based system when combining MFCC with S1 or S2. However, the speaker verification system based on i-vectors + PLDA and using as feature vectors the fusion of MFCC + S1 achieves an EER = 2.11%, which is the lowest EER for all systems we are comparing. These results are also in line with previously reported research in [20] where fusion was used as a strategy to enhance performance in a speaker verification system. Finally, to complement results presented in Tables 2 and 3, Fig. 4 depicts the comparison of all systems using the DET curve [44]. As can be seen, the MFCC+S1 / GMM-UBM based system presents the best performance when compared to other approaches.

5. Conclusions

Herein, we have addressed the problem of speaker verification (SV) using articulatory information extracted from the tongue movements. According to our results, articulatory features proposed in this work contain highly discriminative information associated to speaker identity, which are useful for speaker verification purposes. It is also important to highlight that these features contain complementary information that can be used in a fusion scheme with typical acoustic features, such as MFCC, improving speaker verification performance. Features extracted from the tongue contour using a polynomial approach exhibit promising results, achieving an EER = 3.73%, which is higher to the EER achieved by MFCC alone (2.53%), but still comparable.

Even though articulatory information has been used previously for speaker recognition purposes, this work presents an approach that reduces interference and cost while acquiring the signal based on ultrasound images. We also show the potential of this approach to extract complementary information to typical acoustic features, and this is shown when fusing features in a single vector, which results in a better performance of the speaker verification system.

On the other hand, for the task of estimating articulatory information, it is possible to use speaker independent acoustic-to-articulatory inversion methods. However, these methods have been evaluated in a small amount of speakers for the task at hand. In this regard, ultrasound technology allows to measure articulatory information directly from more speakers and at a reduced cost, thus, obtaining better acoustic-to-articulatory models. These new models could be used as a complement for acoustic features. For future work, we propose to implement an speaker independent acoustic-to-articulatory inversion method for feature extraction to be used in speaker recognition systems.

References

[1] Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V. and Wellekens, C., Automatic speech recognition and speech variability: a review. Speech Communication. 49(10), pp. 763-786, 2007. DOI: 10.1016/j.specom.2007.02.006

[2] O'Shaughnessy, D., Speech communications: human and machine, 2nd Ed., Wiley-IEEE Press, New York, USA, 1999, 548 P.

[3] Kitapci, K. and Galbrun, L., Perceptual analysis of the speech

intelligibility and soundscape of multilingual environments. Applied Acoustics, 151, pp. 124-136, 2019. DOI: 10.1016/j.apacoust.2019.03.001.

[4] Rix, A.W., Beerends, J.G., Hollier, M.P. and Hekstra, A.P., Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs, in: International Conference on Acoustics, Speech, and Signal Processing, Proceedings. IEEE, Salt Lake City, USA, 2001, pp. 749-752.

[5] Kinnunen, T. and Li, H., An overview of text-independent speaker recognition: from features to supervectors. Speech Communication, 52(1), pp. 12-40, 2010. DOI: 10.1016/j.specom.2009.08.009.

[6] Reynolds, D.A., Experimental evaluation of features for robust speaker identification. IEEE Transactions on Speech and Audio Processing, 2(4), pp. 639-643, 1994. DOI: 10.1109/89.326623.

[7] Doddington, G.R., Przybocki, M.A., Martin, A.F. and Reynolds, D.A., The NIST speaker recognition evaluation - Overview, methodology, systems, results, perspective. Speech Communication. 31(2-3), pp. 225-254, June, 2000. DOI: 10.1016/S0167-6393(99)00080-1.

[8] Bishop, C.M., Pattern recognition and machine learning, 1st Ed., Springer, New York, USA, 2006.

[9] Duin, PW. R. and Pekalska, E., Dissimilarity representation for pattern recognition, the: foundations and applications. 1st Ed., World scientific, New Jersey, USA, 2005.

[10] Dehak, N., Kenny, P.J, Dehak, R., Dumouchel, P. and Ouellet, P., Front-end factor analysis for speaker verification. Transactions on Audio, Speech, and Language Processing, 19(4), pp.788-798, 2011. DOI: 10.1109/TASL.2010.2064307.

[11] Kenny, P., Boulianne, G., Ouellet, P., Dumochel, P., Joint Factor Analysis Versus Eigenchannels in Speaker Recognition. In: IEEE Transactions on Audio, Speech, and Language Processing, 15(4), pp. 1435-1447, 2007. DOI: 10.1109/TASL.2006.881693.

[12] Sreenivasa, K. and Sarkar, S., Robust speaker recognition in noisy environments, 1st Ed., Springer, New York, USA, 2014, pp. 2-49.

[13] Leung, K., Mak, M., Siu, M. and Kung, S., Adaptive articulatory feature-based conditional pronunciation modeling for speaker verification. Speech Communication, 48(1), pp. 71-84, 2006. DOI: 10.1016/j.specom.2005.05.013.

[14] Dromey, C. and Sanders, M., Intra-speaker variability in palatometric measures of consonant articulation. Journal of Communication Disorders, 42(6), pp. 397-407, 2009. DOI: 10.1016/j.jcomdis.2009.05.001.

[15] Serruirer, A., Badin, P., Bo, L., Lamalle, L. and Nesuchaefer-Rube, C., Inter-speaker variability: speaker normalisation and quantitative estimation of articulatory invariants in speech production for French, in: Speech and Language Processing, 4th, 2017, Interspeech, Stockholm, Sweden, 2017.

[16] Ghosh, P.K. and Narayanan, S., A generalized smoothness criterion for acoustic-to-articulatory inversion. The Journal of the Acoustical Society of America, 128(4), art. 2172, 2010. DOI: 10.1121/1.3455847.

[17] Sepúlveda, A., Capobianco, R. and Castellanos, G., Estimation of relevant time frequency features using Kendall coefficient for articulator position inference. Speech Communication, 55(1), pp. 99-110, 2013. DOI: 10.1016/J.SPECOM.2012.06.005.

[18] Potard, B., Laprie, Y. and Ouni, S., Incorporation of phonetic constraints in acoustic-to-articulatory inversion. Journal of the Acoustical Society of America, 123(4), pp. 2310-2323, 2008. DOI: 10.1121/1.2885747.

[19] Ghosh, P.K. and Narayanan, S., Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion. The Journal of the Acoustical Society of America, 130(4), pp. EL251-EL257, 2011. DOI: 10.1121/1.3634122.

[20] Li, M., Kim, J., Lammert, A., Ghosh, P.K., Ramanarayanan, V. and Narayanan, S., Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Computer Speech & Language, 36, pp. 196-211, 2016. DOI: 10.1016/j.csl.2015.05.003.

[21] Aravind, I. and Ghosh, P.K., Inferring speaker identity from articulatory motion during speech, in: Workshop on Machine Learning in Speech and Language Processing 5th Interspeech, 2018, Hyderabad, India, 2018.

[22] Aron, M., Kerrien, E., Berger, M. and Laprie, Y., Coupling electromagnetic sensors and ultrasound images for tongue tracking in International Seminar on Speech Production, 7th, ISSP, 2006, Ubatuba-SP, Brazil. 2006.

[23] Narayanan, S., Toutios, A., Ramanarayanan, V., Lammert, A., Kim, J., Lee, S., Nayak, K., Kim, Y., Zhu, Y., Goldstein, L., Byrd, D., Bresch, E., Ghosh, P., Katsamanis, A. and Proctor, M., Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). The Journal of the Acoustical Society of America, 136(3), pp. 1307-1311, 2014. DOI: 10.1121/1.4890284.

[24] Prasad, A., Periyasamy, V. and Ghosh, P., Estimation of the invariant and variant characteristics in speech articulation and its application to speaker identification, in: International Conference on Acoustics, Speech and Signal Processing (ICASSP), 40th, IEEE, 2015, Brisbane, Australia, 2015, pp. 4265-4269.

[25] Porras, D., Sepúlveda, A. and Csapó, G., DNN-based acoustic-to-articulatory inversion using ultrasound tongue imaging, in: 21st, International Join Conference on Neural Networks, IJCNN, Budapest, Hungary, 2019.

[26] Scobbie, J., Wrench, A. and van der Linden, M., Head-probe stabilisation in ultrasound tongue imaging using a headset to permit natural head movement, in: 8th International seminar on speech production. Proceedings of the 8th International seminar on speech production, Strasbourg, France, 2008, pp. 373-376.

[27] Whalen, D., Iskarous, K., Tiede, M., Ostry, D., Lehnert-LeHouillier, H., Vatikiotis-Bateson, E. and Hailey, D., The Haskins optically corrected ultrasound system (HOCUS). Journal of Speech, Language, and Hearing Research, 48(3), pp. 543-553, 2005. DOI: 10.1044/1092-4388(2005/037).

[28] Castillo, M., Rubio, F., Porras, D., Contreras, S. and Sepúlveda, A., A small vocabulary database of ultrasound image sequences of vocal tract dynamics, in: 2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA) Conference Proceedings, IEEE, Bucaramanga, Colombia, 2019, pp. 1-5.

[29] Li, M., Kambhamettu, C. and Stone, M., Automatic contour tracking in ultrasound images. Clinical Linguistics & Phonetics, 19(6-7), pp. 545-554, 2005. DOI: 10.1080/02699200500113616.

[30] Kass, M., Witkin, A. and Terzopoulos, D., Snakes: active contour models. International Journal of Computer Vision, 1(4), pp. 321-331, 1988. DOI: 10.1007/BF00133570.

[31] Proakism, J. and Manolakis, D., Digital signal processing: principles, algorithms, and applications, 3rd Ed, Prentice Hall, N.J., United States of America, 1996.

[32] Childers, D.G., Skinner, D.P. and Kemerait, R.C., The cepstrum: a guide to processing. Proceedings of the IEEE, 65(10), pp. 1428-1443, 1977. DOI: 10.1109/PROC.1977.10747.

[33] O'Shaughnessy, D., Invited paper: automatic speech recognition: history, methods and challenges. Pattern Recognition, 41(10), pp. 2965-2979, 2008. DOI: 10.1016/j.patcog.2008.05.008.

[34] Ichikawa, O., Fukuda, T. and Nishimura, M., Dynamic features in the linear-logarithmic hybrid domain for automatic speech recognition in a reverberant environment. IEEE Journal of Selected Topics in Signal Processing, 4(5), pp. 816-823, 2010. DOI: 10.1109/JSTSP.2010.2057191.

[35] Zheng, F., Zhang, G. and Song, Z., Comparison of different implementations of MFCC, Journal of Computer Science and Technology, 16(6), pp. 582-589, 2001.

[36] Young, S., Evermann, G., Kershaw, D., Moore, G., Gales, M., Odell, J., Hain, T., Liu, X., Ollason, D., Povey, D., Valtchev, V. and Woodland, P. The htk book [online], version 3.2, Cambridge University Engineering Department, Cambridge, UK, 2002. [Consulted, September 14th of 2018]. Available at: http://www.dsic.upv.es/docs/posgrado/20/RES/materialesDocentes/alejandroViewgraphs/htkbook.pdf.

[37] Brookes, M., Voicebox: speech processing toolbox for matlab. [online]. 2nd Ed., Department of Electrical & Electronic Engineering, Imperial College, London, UK, 2011. [Consulted, September 20, 2018]. Available at: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[38] Sadjadi, S.O., Slaney, M. and Heck, L., Msr identity toolbox v1.0: a Matlab toolbox for speaker-recognition research. Speech and

Language Processing Technical Committee Newsletter, 1(4), pp.1-32, 2013.
[39] Reynolds, D.A., Quatieri, T.F. and Dunn, R.B., Speaker verification using adapted gaussian mixture models. Digital Signal Processing, 10(1-3), pp.19-41, 2000. DOI: 10.1006/dspr.1999.0361.
[40] Bousquet, P-M., Matrouf, D. and Bonastre, J-F., Intersession compensation and scoring methods in the i-vectors space for speaker recognition, in: 12th Annual Conference of the International Speech Communication Association Proceedings, ISCA, Florence, Italy, 2011, pp. 485-488.
[41] Sizov, A., Lee, K.A. and Kinnunen, T., Unifying probabilistic linear discriminant analysis variants in biometric authentication, in: 10th The Joint Biannual Event Statistical Pattern Recognition Technique and Structural and Syntactical Pattern Recognition, Proc. S+SSPR, IAPR, Joensuu, Finland, 2014, pp. 464-475.
[42] Simon, J.D., Prince and Elder, J.H., Probabilistic linear discriminant analysis for inferences about identity, in: 11th International Conference on Computer Vision, Proceedings ICCV, IEEE, Rio de Janeiro, Brazil, 2007, pp. 1-8.
[43] Zhang, Y., Long, Y., Shen, X., Wei, H., Yang, M., Ye, H. and Mao, H., Articulatory movement features for short-duration text-dependent speaker verification. International Journal of Speech Technology, 20(4), pp. 753-759, 2017. DOI: 10.1007/s10772-017-9447-8.
[44] Martin, A., Doddington, G., Kamm, T., Ordowski, M. and Przybocki, M., The det curve in assessment of detection task performance, Gaithersburg, USA, National Institute of Standards and Technology, 1997.

D. Porras-Plata, received the BSc. of Electronic Engineering in 2016, and MSc. in Electronic Engineering in 2019, all of them from the Universidad Industrial de Santander, Bucaramanga, Colombia. Since 2015 he has worked researching in speech signal processing. In 2018 he completed a research internship at the SMARTLAB laboratory, at the University of Technology and Economics (BME), Budapest, Hungary. Where he implemented a DNN, to perform an acoustic-articulation mapping. His research interests include: deep learning, machine learning.
ORCID: 0000-0002-2254-0513

A. Sepulveda-Sepulveda, is a BSc. Eng. an Electronic Engineer and MSc. in Industrial Automation from the Universidad Nacional de Colombia, in 2004 and 2006, respectively. Later he received the same university degree in Dr. in 2012. He is currently an associate professor at the School of Electrical, Electronic and Telecommunications Engineering of the Universidad Industrial de Santander, Bucaramanga, Colombia. His areas of interest correspond to the treatment of speech signals and applications of machine learning.
ORCID: 0000-0002-9643-5193

M. Sarria-Paja, is BSc. Eng. an Electronic Engineer and MSc. in Industrial Automation from the Universidad Nacional de Colombia, in 2006 and 2009, respectively. He then received a PhD in 2017 from the Institut National de la Recherche Scientifique (INRS-EMT) University of Quebec, Montreal, QC, Canada. He currently works as a full-time professor for the Faculty of Engineering of the Universidad Santiago de Cali, Colombia. His areas of interest correspond to artificial intelligence using machine learning and signal processing, and its relationship with mathematics and statistics.
ORCID: 0000-0003-4288-1742