SecuVoice: A Spanish Speech Corpus for Secure Applications with Smartphones

Juan M. Martín-Doñas[†], Iván López-Espejo[⋆], Carlos R. González-Lao[⋆], David Gallardo-Jiménez[†], Angel M. Gomez[⋆], José L. Pérez-Córdoba[⋆], Victoria Sánchez[⋆], Juan A. Morales-Cordovilla[⋆], and Antonio M. Peinado[⋆]

Dept. of Signal Theory, Telematics and Communications,
University of Granada, Spain
[†]{mdjuamart,davidgj94}@correo.ugr.es
[⋆]{iloes,clao,amgg,jlpc,victoria,jamc,amp}@ugr.es

Abstract. In this paper, a new speech database, the so-called SecuVoice, is described. This database consists of utterances in Spanish of isolated digits recorded with two different smartphones: a mid-range smartphone and a high-range one. This database is intended for research on biometrics and secure applications that integrate both automatic speech recognition (ASR) and speaker recognition/verification. In this regard, both ASR and speaker verification baselines are given in this paper as reference. The experimental results show that a very high performance can be obtained on this corpus. SecuVoice will be released through ELRA (European Language Resource Association), so that speech researchers can evaluate and compare the performance of their speech-related developments and algorithms within a framework with speech signals acquired with real smartphones.

Keywords: SecuVoice, Speech database, Smartphone, Secure application

1 Introduction

Smartphones have become an essential tool in our society. These devices not only allow us to make phone calls, but a large number of additional tasks. Many of them can be performed in a natural and comfortable way for the user by means of her/his voice. Some examples of speech-related tasks that can be carried out with smartphones are search-by-voice, dictation or telebanking. Of course, signal processing and, more particularly, speech/audio processing techniques are involved in such applications. For instance, speech enhancement algorithms can be used to improve the speech quality perceived by the speakers during a phone call [1, 2]. Also, automatic speech recognition (ASR) is necessary for search-by-voice, dictation applications and access granting through a temporary key

[⋆] This work has been supported by the Spanish MINECO TEC2013-46690-P project and the INNPACTO project "SecuVoice: Voice Biometrics to Guarantee the Security of Enterprise Applications" (IPT-2012-0082-390000).

[3]. Likewise, to provide security in telebanking operations applying speaker recognition techniques could be required in order to verify the identity of a customer [4].

To guarantee a good user experience when running that kind of tools, it is of utmost importance (especially for a secure environment) to design and develop robust and high-quality speech processing methods. These methods must show their worth in validation tests, which requires the use of statistically representative speech databases.

In this paper we describe a new speech database called SecuVoice recorded as a development and test tool for the INNPACTO project "SecuVoice: Voice Biometrics to Guarantee the Security of Enterprise Applications" (IPT-2012-0082-390000). This database of spoken Spanish is comprised of utterances of isolated digits recorded in an office environment with a mid-range and a high-range smartphone. Moreover, due to the structure of SecuVoice, which will be presented in the next sections, this database is especially suitable for research on biometrics and secure applications that integrate both ASR and speaker recognition/verification. In this regard, the utterances of this corpus might be considered as one time passwords (OTP) uttered by the users by employing their smartphones within a context of a remote secure system. Such a system might apply ASR to check that the user knew and uttered the OTP correctly as well as a speaker recognition/verification procedure in order to authenticate her/him. SecuVoice will be available through ELRA (European Language Resource Association) [5] to any researcher/individual interested in this language resource.

The rest of this paper is organized as follows. First, in Section 2, the speech recording procedure carried out per speaker is explained along with the recording acoustic conditions. Information about the speakers that participated in the corpora recording is presented in Section 3. In Section 4 we describe how the speech data are arranged into datasets as well as we explain the content of the annotation files provided with detailed information about both the speakers and the recordings. Section 5 is devoted to explain the frameworks considered in the development of the speech recognition and speaker verification systems and the results obtained from their evaluation using the database. Finally, some conclusions are summarized in Section 6.

2 Data Recording

SecuVoice's corpora consists of single-channel utterances in Spanish containing sequences of isolated digits from *zero* (*cero*) to *nine* (*nueve*). These utterances were acquired in spring 2013 by using two different devices, i.e. a mid-range smartphone and a high-range one. The mid-range smartphone used was an HTC WildFire while the high-range device was a Sony Xperia S. For both models, the utterances were stored as uncompressed monophonic WAV files with a sampling frequency of 8000 Hz and 16 bits per sample.

Utterance	1st session	2nd session	3rd session
Enrollment	2074539681	4179536280	5314986072
1st verif.	0142	1437	1005
2nd verif.	8937	5698	3178
3rd verif.	5669	3170	6924
4th verif.	0487	4526	8215
5th verif.	5321	3645	0937
6th verif.	8920	2798	4635

Table 1. Sequences of digits uttered by every speaker per smartphone.

The voice of every speaker was recorded over three sessions, lasting around ten minutes each. Furthermore, in order to ensure that the acquired speech samples are representative, a gap of at least two weeks was left between two consecutive sessions. The recording protocol was exactly the same in each session and consisted of the following steps for every phone model:

1. The speaker was given one smartphone (usually the HTC WildFire in first instance).
2. The device was held in one hand by the speaker at a certain distance from her/his face in order to read the digits that an application running on the smartphone displayed on its screen.
3. The application generated on the screen of the device a sequence of ten digits (from *zero* to *nine* in a fixed randomized order). That sequence was uttered by the speaker and recorded by the application. The resulting utterance is known as *enrollment* utterance. It should be noted that the application approximately generates one digit per second, forcing a brief pause between each two consecutive digits in an utterance.
4. Similarly to the case of the *enrollment* utterance, six sequences of four digits each were uttered by the speaker and recorded by the application. The resulting utterances are known as *verification* utterances.

Finally, the procedure described above was repeated with the second device (usually the Sony Xperia S). Thus, at the end of each session 14 utterances were recorded per speaker with a total of 68 digits. Therefore, at the end of the three sessions every speaker contributed to the SecuVoice's corpora with 42 utterances and a total of 204 digits (i.e. 21 utterances and 102 digits per smartphone model). It must be remarked that every speaker uttered the same set of randomized sequences of digits per smartphone, which are shown in Table 1. As can be seen, every digit from *zero* to *nine* appears 10 times each one, excepting the digits *three* and *five* with 11 occurrences each one.

It must be noted the following considerations about the data recording procedure. First of all, the speakers were encouraged to properly vocalize as well as change their intonation and rhythm over each sequence of digits. Also, within

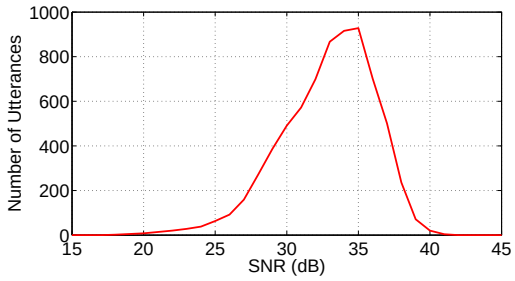

Fig. 1. Histogram of estimated SNRs of the complete SecuVoice's corpora (one SNR value per utterance).

reasonable limits speakers were able to choose a comfortable distance between themselves and the smartphone during the reading of the digits. In fact, some speakers slightly changed that distance while reading the digits. The speakers were also free to choose the volume of their voice, except when it was too low. In the latter case the speaker was encouraged to raise her/his voice. Finally, we verified that all the sequences of digits were read correctly.

2.1 Acoustic Environment

An office environment was considered to obtain the SecuVoice's corpora. Thus, all the speech recordings were made in a rather silent room, the size of which is about 12 m^2, with some office furniture: four office chairs, two desks, shelves and a round table in the middle. It should be noted that the speakers sat at that table when recording the digits.

Although the ambient noise during the speech acquisition procedure was low, we can enumerate a range of noises which may have slightly affected the recordings. These noises are the following: noise from the cooling system in the room, noise from a laptop, little bangs on the table accidentally made by the speaker, noise from the chair where the speaker was sitting, slamming doors in adjacent rooms and babble noise from nearby rooms and corridors. In general, the amplitude of such noises is quite low and their duration very short. Indeed, we must highlight that we tried to avoid those ambient noises as much as possible in such a way that they are nearly absent in the recordings. This is confirmed by Figure 1, where a histogram with the number of utterances per estimated signal-to-noise ratio (SNR) in SecuVoice is shown. The average SNR is 32.9 dB.

3 About the Speakers

A total of 169 adult speakers participated in the recording of the SecuVoice's corpora. 128 of them were male while 41 were female. Furthermore, although we can find in this corpus speakers from the age of 18 years to the age of 50

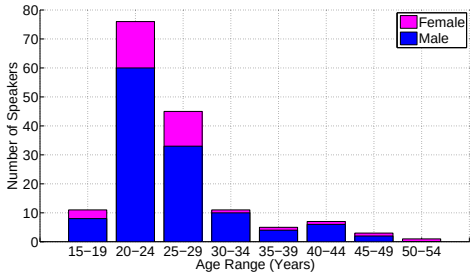

Fig. 2. Number of male and female speakers per age range in SecuVoice's corpora.

years, most of the speakers are in the range from 20 years to 30 years as can be observed in Figure 2.

We can highlight the richness of the database in terms of the variety of both the accents present in it and the geographical origin of the speakers. On the one hand, most of the speakers (127 of 169) are from Eastern Andalusia. In a lesser proportion (34), we can find some examples from the rest of the Iberian Peninsula and the Canary Islands as well as from some Latin American countries. Similarly, some people whose mother tongue is not Spanish (8) participated in these recordings. In this regard, there are representatives from the following countries: Algeria, Belgium, Brazil, England, France, India, Italy and Philippines.

4 Structure of the Database

SecuVoice's corpus is comprised of a total of 169 speakers $\times$ 42 utt./speaker = 7098 utterances with a total of 34476 digits (204 digits/speaker). Each digit from *zero* to *nine* is present 3380 times excepting the digits *three* and *five* with a number of occurrences of 3718 each one. Utterances are arranged into two different datasets, i.e. the *enrollment* (ENROLL) and *verification* (VERIF) datasets. The ENROLL dataset is composed of the 1014 *enrollment* utterances (169 speakers $\times$ 6 enroll. utt./speaker) with 10140 digits, where the digits from *zero* to *nine* are balanced. On the other hand, the VERIF dataset consists of the 6084 *verification* utterances (169 speakers $\times$ 36 verif. utt./speaker) with 24336 digits. In the latter dataset, each digit from *zero* to *nine* is present 2366 times excepting the digits *three* and *five*, which appear 2704 times each one.

In SecuVoice, along with the WAV files containing the speech utterances, annotation files based on the XML (eXtensible Markup Language) [6] format are provided. There are two types of annotation files containing detailed information about the speakers and the recorded sequences of digits, i.e. the speaker annotation file (one per speaker) and the utterance annotation file (one per WAV file). In the following subsections these types of files are described.

4.1 The Speaker Annotation File

The speaker annotation file contains all the information describing the main characteristics of a speaker. The fields that we can find in this XML file are the following:

- *IdSpeaker*: This is a unique identifier assigned to each speaker. The possible values for this identifier are in the range from 30001 to 30200.
- *Gender*: This identifies the gender of the speaker, *M* if male or *F* if female.
- *Age*: The age of the speaker is specified in this field.
- *Session*`X`: This contains some information related to the 1st (`X`=1), 2nd (`X`=2) or 3rd (`X`=3) speaker recording session. In particular, both the date (*Date*) and time (*Time*) of the session are noted. Dates appear in the `DD/MM` format (it should be reminded that the SecuVoice's corpora was recorded within 2013). *Time* is set to *M* if the session was between 10 a.m. and 2 p.m. or *A* if it was between 4 p.m. and 8 p.m.
- *Charact*: This field shows qualitative information about the speaker as well as about the characteristics of her/his voice and speech. Depending on the speaker, it can be found some information about her/his accent, geographical origin, pronunciation and diction, intonation and rhythm, volume of voice, tone and pitch, etc.

4.2 The Utterance Annotation File

Each WAV file containing the recording of a sequence of digits has a corresponding utterance annotation file. This file provides all the necessary information about the utterance, and we can find the following fields in it:

- *IdSpeaker*: The aforementioned identifier that references the speaker that uttered the sequence of digits is shown in this field.
- *IdSession*: This is the session in which the utterance was recorded. The value of this field is 1, 2 or 3 if the corresponding recording session was the first, second or third one, respectively.
- *Device*: This identifies the device used to record the utterance. The value of this field is *MID* if the HTC WildFire was used or *HIGH* if the employed smartphone was instead the Sony Xperia S.
- *TypeSequence*: If the utterance belongs to the *enrollment* dataset, *TypeSequence* is set to *ENROLL*. On the other hand, the value of this field is *VERIFY* if the utterance is from the *verification* dataset.
- *Digits*: This is the transcription of the sequence of digits present in the utterance.
- *digit*`X`: This contains the information needed to segment a particular digit out of the utterance, where `X` indicates its position within the sequence. For example, let us consider an utterance with the sequence of digits *one five two nine*. If we are interested in the segmentation information of the digit *two*, we will look for the tag *digit3* (i.e. `X`=3). In turn, labels *start_digit* and *end_digit* identify the initial and ending samples, respectively, which delimit

the digit within the WAV speech file. It must be noted that the pauses immediately after and before the digit are included within this delimitation. Sometimes, a small noise appears over the pause periods between digits. For these cases, a second tight segmentation is provided where part or all of the pause immediately before and immediately after the digit is removed. In this case, labels *start_tight_digit* and *end_tight_digit* indicate the digit initial and ending samples. Furthermore, it should be noticed that this second type of segmentation is not always given. In every case, the digits were manually segmented.

– *Incidences*: Relevant aspects observed during the recording are noted in this field, e.g. the appearance of some noise from those referenced in Subsection 2.1.

5 Database Performance Evaluation

In this section we show the results obtained when evaluating the database for speech recognition and speaker verification purposes. In the former case the objective is to train an ASR system to be able to recognize the digits that each speaker has spoken. In the latter case we look for a system which verifies the identity of a speaker in order to avoid impostors. The following subsections are devoted to present both the frameworks that we have used for these two tasks and the results from the evaluation of the trained systems.

5.1 Speech Recognition Results

For the speech recognition task we use the HTK toolkit [7]. To evaluate our ASR system we consider isolated digits. Therefore, SecuVoice utterances are segmented so new utterances are created with only one digit each, including initial and final silence. The information contained in the utterance annotation files (*start_digit* and *end_digit* fields) is used to this end.

The acoustic features are extracted from the segmented speech signals using the European Telecommunication Standards Institute front-end (ETSI FE, ES 201 108) reported in [8]. Twelve Mel-frequency cepstral coefficients (MFCCs) along with the 0th order coefficient and their respective velocity and acceleration components form the 39-dimensional feature vector used by the recognizer. The speakers are divided in two subsets: A and B. The first 100 speakers are grouped in subset A while the remaining 69 speakers in subset B. The segmented *enrollment* and *verification* utterances from the speakers in subset A are used to train the acoustic models. Left to right continuous density hidden Markov models (HMMs) with 16 states and 8 Gaussians per state are used to model each digit. Silences are modeled by HMMs with 3 states and 16 Gaussians per state.

For the evaluation of the recognition accuracy the following three approaches are considered. The first one uses the *verification* utterances from subset B for testing, yielding a word accuracy of **99.67%**. The second approach also uses

the VERIF dataset from subset B for testing but cepstral mean and variance normalization (CMVN) is applied to the whole set of training and testing speech features. The word accuracy obtained in this case is **99.62%**. Finally, the last approach is as the second one but the *enrollment* utterances from subset B are used to perform speaker adaptive training using maximum likelihood linear transformation (MLLR), as described in [7]. Hence, the transformed models for each speaker are used to evaluate the *verification* utterances from subset B, yielding a word accuracy of **99.84%**.

5.2 Speaker Verification Results

The front-end for the speaker verification task is composed of the following stages: voice activity detection to remove the silence segments, pre-emphasis filtering, extraction of MFCC features (14 coefficients along with their respective velocity and acceleration) and CMVN.

In order to carry out a performance evaluation, a jackknife-like test is applied. To do this, the whole database (169 speakers) is segmented into 13 blocks with 13 speakers each. For every jackknife iteration, a subset of 7 blocks (91 speakers, *enrollment* + *verification* sentences) is employed to train a 256-component universal background model (UBM), while the remaining blocks are reserved for testing: 3 blocks (39 speakers) as granted speakers and 3 blocks (39 speakers) as impostors. In the first iteration, blocks 1-7 are used for UBM training, while blocks 8-10 are used as granted speakers and blocks 11-13 for impostors. In the second iteration, blocks 8-10 and 11-13 are exchanged as granted speakers and impostors. The third and fourth iterations employ blocks 2-8 for UBM training, and (9,10,11) and (12,13,1) for granted speakers and impostors. Every two iterations a 1-block shift is applied to generate the required subsets (i.e. UBM training subset, granted speakers and impostors) until a circular series of 13 shifts is completed. This results in a total of $13 \times 2 = 26$ jackknife iterations.

In each of these jackknife iterations, the *enrollment* utterances corresponding to the subset of granted speakers are used to obtain the model's $\mathbf{T}$ matrix, which defines the total variability subspace. Then, an i-vector is extracted per *enrollment* utterance. A linear discriminant analysis (LDA) is applied to reduce their dimensionality to 38 components and, then, a Gaussian probabilistic LDA (G-PLDA) model is obtained by using these i-vectors. Finally, every speaker's i-vector is obtained from the mean of the 6 i-vectors from her/his corresponding *enrollment* utterances. More details about speaker model computation can be found in [9, 10].

During evaluation, the 36 *verification* utterances of every testing speaker (at every jackknife iteration) are employed. Every testing utterance is processed as mentioned above in order to obtain its corresponding i-vector. For false negative rate estimation, the 36 *verification* utterances of every granted speaker are matched against its corresponding speaker model. Similarly, the false positive rate can be obtained by matching the 36 *verification* utterances of every impostor against the 39 available speaker models. As a result of the whole jackknife

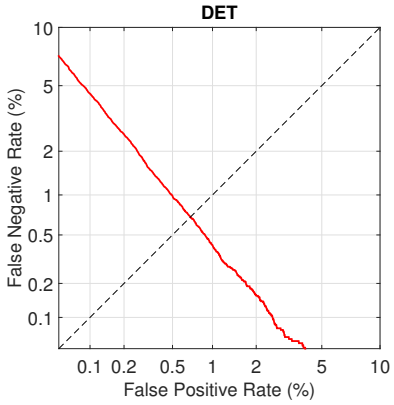

Fig. 3. Detection error trade-off (DET) graph of the evaluated speaker verification system.

Parameter	EER	minDCF (NIST 2008)	minDCF (NIST 2010)
Value	0.69%	0.45%	0.12%

Table 2. Speaker verification system performance.

process, we have that all the 169 speakers have been used as both granted speakers and impostors, and the global false negative and false positive rates can be computed. This is carried out for 1000 different thresholds, obtaining a detection error trade-off (DET) plot with 1000 points as shown in Figure 3. Additionally, the corresponding values of equal error rate (EER) and minimum detection cost function (minDCF, both the NIST 2008 and NIST 2010 approaches) [11] are presented in Table 2.

6 Conclusions

In this paper we have described a new speech database of isolated digits in Spanish, the so-called SecuVoice. This database is intended for research and development on biometrics and secure applications based both on ASR and speaker recognition/verification. Thus, speech researchers can evaluate and compare the performance of their speech-related developments and algorithms within a framework with speech signals acquired with real smartphones. Both speech recognition and speaker verification systems have been developed and evaluated using the database in order to serve as a baseline for future works. Our experimental results have shown that a very high performance is obtained on this corpus. SecuVoice will soon be released through ELRA [5].

References

1. Premananda, B.S., Uma, B.V.: Speech Enhancement to Overcome the Effect of Near-End Noise in Mobile Phones Using Psychoacoustics. In: ICCCNT, pp. 1–6, Hefei, China (2014)
2. Hu, J., Lee, M.: Speech Enhancement for Mobile Phones Based on the Imparity of Two-Microphone Signals. In: ICIA, pp. 606–611, Zhuhai, Macau (2009)
3. Acero, A., Bernstein, N., Chambers, R., Ju, Y.C., Li, X., Odell, J., Nguyen, P., Scholz, O., Zweig, G.: Live Search for Mobile: Web Services by Voice on the Cellphone. In: ICASSP, pp. 5256–5259, Las Vegas, USA (2008)
4. Selvan, K., Joseph, A., Babu, K.K.A.: Speaker Recognition System for Security Applications. In: RAICS, pp. 26–30, Trivandrum, India (2013)
5. European Language Resources Association, http://www.elra.info/en/about/
6. Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E., Yergeau, F.: Extensible Markup Language (XML) 1.0 (Fifth Edition), http://www.w3.org/TR/REC-xml/ (2008)
7. Young, S., et al.: The HTK Book, Version 3.4. Cambridge University Engineering Department (2006)
8. ETSI ES 201 108 - Distributed Speech Recognition; Front-End Feature Extraction Algorithm; Compression Algorithms
9. Kenny, P.: A Small Footprint i-Vector Extractor. In: ISCA Odyssey, The Speaker and Language Recognition Workshop, Singapore (2012)
10. Kenny, P.: Bayesian Speaker Verification with Heavy-Tailed Priors. In: ISCA Odyssey, The Speaker and Language Recognition Workshop, Brno, Czech Republic (2010)
11. Van Leeuwen, D. A., Brümmer, N.: An Introduction to Application-Independent Evaluation of Speaker Recognition Systems. *Lecture Notes in Computer Science*, vol. 4343 (2007)