SecuVoice: A Spanish Speech Corpus for Secure Applications with Smartphones

J. M. Martín-Doñas, I. López-Espejo, C. R. González-Lao, D. Gallardo-Jiménez, A. M. Gomez, J. L. Pérez-Córdoba, V. Sánchez, J. A. Morales-Cordovilla, and A. M. Peinado
Dept. of Signal Theory, Telematics and Communications, University of Granada, Spain

Introduction

What is it?

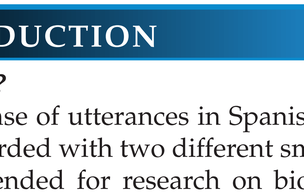
- A database of utterances in Spanish of isolated digits recorded with two different smartphones.
- It is intended for research on biometrics and secure applications that integrate both ASR and speaker recognition/verification.

Motivation

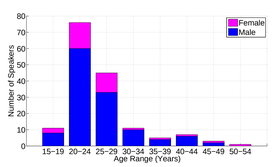
- The importance of speech-related tasks in smartphones and remote secure systems.

Data Recording

- Single-channel utterances of **isolated digits** (0 to 9).
- **Two smartphones**: high-range (Sony Xperia S) and mid-range (HTC WildFire).
- Three sessions of ten minutes each. Far-talk conditions.
- An *enrollment* utterance (10 digits) and six *verification* utterances (4 digits) per session and smartphone.
- Speech recording in a rather silent office of 12 m^2. Average SNR of 32.9 dB.



- A total of **169 speakers**, most of them from Eastern Andalusia.



Gender/age histogram of the SecuVoice's speakers.

Structure of the Database

- Corpus of **42 utterances/speaker** (in all, 7098 utterances).
- Two different datasets: ENROLL and VERIF (with the *enrollment* and *verification* utterances, respectively).
- WAV files containing the speech utterances.
- **XML annotation files**: speaker annotation files (one per speaker) and utterance annotation files (one per WAV file).

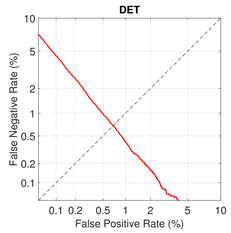
Utterance	1st session	2nd session	3rd session
Enrollment	2074539681	4179536280	5314986072
1st verif.	0142	1437	1005
2nd verif.	8937	5698	3178
3rd verif.	5669	3170	6924
4th verif.	0487	4526	8215
5th verif.	5321	3645	0937
6th verif.	8920	2798	4635

Evaluation: Speaker Verification

Experimental framework

- Front-end: VAD, pre-emphasis filtering, MFCC and CMVN.
- **Jackknife-based test**: 13 blocks (13 speakers each). A total of 26 iterations.
- At each iteration, the 13 blocks are divided as follows: 7 blocks to train a UBM, 3 as granted speakers and 3 as impostors.
- ENROLL dataset of granted speakers used to obtain i-vectors and G-PLDA models.
- VERIF dataset of granted and impostor speakers for testing the system (false positive and negative rates).

Results



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

Evaluation: Speech Recognition

Experimental framework

- Utterances are segmented to consider isolated digits.
- We use the **ETSI front-end** to extract MFCC features.
- Speakers are divided in two subsets: A (100 speakers) and B (69 speakers).
- Subset A (both ENROLL and VERIF datasets) is used to train **GMM-HMM** acoustic models.

Three methods for evaluating the recognition accuracy:

1. VERIF dataset of subset B is used for testing.
2. Same as above, but CMVN is applied to both training and testing features.
3. Same as above, but ENROLL dataset of subset B is used to perform speaker adaptive training by means of MLLR.

Results

Method	*WAcc (%)*
MFCC	99.67
MFCC+CMVN	99.62
MFCC+CMVN+MLLR	**99.84**

Conclusion

- The SecuVoice database has been described, and both speech recognition and speaker verification results are given as baseline. Speech researchers can evaluate and compare the performance of their own algorithms within this framework.
- SecuVoice is available through **ELRA** (European Language Resources Association). http://catalog.elra.info/product_info.php?products_id=1293

Contact Information

Juan M. Martín-Doñas, I. López-Espejo, A. M. Gomez, A. M. Peinado
E-mail: mdjuamart@correo.ugr.es, iloes@ugr.es, amgg@ugr.es, amp@ugr.es
Dept. of Signal Theory, Telematics and Communications, University of Granada, Spain