Building a Basque/Spanish bilingual database for speaker verification

Iker Luengo, Eva Navas, Iñaki Sainz, Ibon Saratxaga, Jon Sanchez, Igor Odriozola, Juan J. Igarza and Inma Hernaez
AhoLab Signal Processing Group, Department of Electronics and Telecommunications,
University of the Basque Country (UPV/EHU)
Alda. Urquijo s/n, 48013 Bilbao
E-mail: {ikerl, eva, inaki, ibon, ion, igor, jigarza, inma}@ aholab.ehu.es

Abstract

Research groups aiming to record new speech databases for minority languages have to face a series of difficulties, such as the lack of previous resources, the scarcity of fluent speakers and the shortage of funding for the project. Sometimes it is possible to take advantage of recording campaigns for other projects, and extend them in order to make some recordings in that language for every contributor that is able to speak it. In this way, new databases can be recorded with little extra effort, as the campaign is already prepared and funded. Using this technique, a new Basque/Spanish bilingual database has been created. Thanks to this database, some research is being undertaken on Basque/Spanish bilingual speaker verification systems. We present a description of the resulting database and the difficulties encountered during its acquisition.

1. Introduction

Currently many systems need some kind of user authentication procedure, in order to verify the users' identity. Most of them use password-type authentication, but passwords may be forgotten or stolen. Nowadays biometric authentication is the best alternative because it can provide extremely accurate and secure access to information (Jain et al., 2006). Furthermore, biometric characteristics cannot be lost nor forgotten and are very difficult to imitate. This kind of authentication can already be seen in different applications such as laptops with fingerprint controlled access and hand geometry based access to certain buildings. Speech is a biometric feature that is non intrusive, has a high degree of acceptability among users and is suitable for long distance verification over data and voice networks. For the development of these speech-based authentication systems, speech databases containing recordings from different speakers are needed.

As a biometric authentication method, a speaker verification system has to decide whether or not a person is who she or he claims to be, using one or more spoken utterances produced by this person (Campbell, 1997). In a generic speaker verification system two modules can be distinguished: the enrolment or training module (which produces a model of every user of the system) and the verification or test module (which decides whether a test utterance has been spoken by an specific speaker)(Naik, 1990)(Bimbot et al., 2004). Usually the language of the training and testing utterances is forced to be the same. But in some environments where bilingual speakers are common, it is desirable that the users of the speaker verification application may be able to utilise any of their known languages to address the system. Therefore, in the last few years various researchers have focused their attention on speaker recognition systems in multilingual environments, where speaker models may be trained with recordings in one language but testing is performed with utterances spoken in another language (Nordstrom et al., 1998)(Faundez-Zanuy & Satue-Villar, 2006). This multilingual environment involves some additional difficulties for the verification system. On the one hand language mismatch between enrolment and verification data produces a degradation in the results of the speaker verification system (Ma & Meng, 2004). On the other hand language mismatches between the target speaker and the world model in a GMM speaker verification system make its performance worse (Auckenthaler et al., 2001).

This bilingual environment is found in the Basque Country. The Basque Country extends north and south of the western side of the Spanish-French border. In the Basque Autonomous Community, situated in the south of the Basque Country, both Basque and Spanish are spoken. Basque is a minority language and therefore there is a scarcity of linguistic resources in this language (Diaz de Illarraza et al., 2003). Specifically, there is no public speech database in Basque available for the development of speaker verification systems.

This paper presents the work and difficulties of recording a new bilingual speech database in the Basque Country for the development of bilingual speaker verification systems in both Spanish and Basque. Section 2 analyses the problems associated with the recording of new speech databases for minority languages. Section 3 describes the recorded bilingual database and its contents, while in section 4 the difficulties that arose during its acquisition are described. Finally, some conclusions are extracted in section 5.

2. Dealing with minority languages in the acquisition of speech databases

Although they are of great social interest, minority and endangered languages are usually not economically interesting, i.e. as there are very few speakers using those languages, it is not worth investing big amounts of money

for research and development of new resources, let them be new corpora compilations or speech databases. This means that finding funding resources for these projects is usually difficult and very often this funding is short in the case of getting any. Moreover, typically there are not many research groups working in these languages, and frequently only one or two groups have interest in a specific language. So collaborative work among research groups in order to distribute work load and expenses is difficult too.

In the case of speaker verification databases, some of the requirements make the process even harder. On the one hand, the recordings in these databases should span along a time period long enough to collect the intra-speaker variation of the voice (Kenny & Dumouchel, 2004). This means that each speaker should be recorded more than once, in different time-spaced sessions, which makes the recording process longer and more expensive. On the other hand, it is desirable that the age and gender distributions of the speakers in the database approximately match the distribution of the expected users. This restriction, together with the fact that being a minority language there may be few speakers available, makes the recruitment of contributors more difficult.

In order to make the acquisition of new databases for minority languages feasible, it can be interesting to take advantage of recording campaigns organized for other projects and use the opportunity to make additional recordings in the minority language too, even though that is not the main objective of the campaign. In this way, new databases can be obtained with little extra effort, as the whole campaign is already prepared.

In the AhoLab Signal Processing Laboratory of the University of the Basque Country research on speech technologies for Basque is being carried out, mainly related to text to speech (TTS), automatic speech recognition (ASR) and speaker recognition. For the development of this research, speech databases in Basque are needed. Sometimes, when there is a database recording campaign for other projects (mainly Spanish speech recordings), contributors are asked to make some extra recordings in Basque, in order to complete a parallel Basque database.

3. Description of the database

The new Basque/Spanish bilingual speaker database was recorded together with a multimodal biometric database acquired in five different Universities all along Spain, including the University of the Basque Country (Galbally et al., 2007). In this database different biometric features were acquired, like fingerprints, signature, handwriting, images of the iris and speech (in Spanish). Seizing the opportunity, the contributors to the database in the University of the Basque Country that were fluent in Basque were also recorded in this language. In this way a small bilingual speaker verification database could be built with little extra effort.

3.1 Design of the database

The recording protocol included four sessions distributed along the time to ensure the capture of the intra-speaker variations that arise as time goes by. There is a difference of two weeks between the recording of the first and second sessions, four weeks between the second and third sessions and six weeks between the third and fourth sessions.

The content of the recordings in each session is:

- Session 1: 4 isolated sentences and 4+3 numeric sequences
- Session 2: 2 isolated sentences and 4+3 numeric sequences
- Session 3: 2 isolated sentences and 4+3 numeric sequences
- Session 4: 2 isolated sentences and 4+3 numeric sequences

The numeric sequences are formed by 8 digits that the speaker reads as she or he prefers. Every speaker has a unique numeric sequence that she or he repeats four times in each session. In addition, she or he records the numeric sequence assigned to three other speakers, which are different in each session, in order to be used as impostor trials. All numeric sequences recorded in each session are common for Spanish and Basque.

The isolated sentences are phonetically rich and balanced, i. e., the distribution of phones in the sentences is similar to the one found in the language. These sentences are the same for all the speakers and are changed from session to session. Obviously, the sentences are different for Spanish and Basque. Therefore the recorded corpus includes 10 different phonetically rich sentences for Spanish and 10 for Basque. In order to select the most appropriate sentences a big corpus used as a reference of the language must be compiled. Once this initial corpus for each language was collected and analysed, the sentences were selected using a software tool called CorpusCrt, made by the TALP research group from the UPC [1], which produces a reduced set of sentences keeping the original frequency of the phonemes as far as it is possible.

The recordings were made in the half-silent environment of a research laboratory, using a Plantronics DSP-400 headset microphone. They were sampled at 44.1KHz and quantified using 16 bits per sample.

3.2 Additional data

The recordings in the database were further processed in order to extract some additional information, namely voice activity and pitch curves.

Voice activity estimation is necessary in order to reject those frames in which there is no vocal information. In this way, noise level during speech silences will not corrupt the features calculated for the speaker verification system. A voice activity detector (VAD) was implemented, based on the computation of the long term spectral deviation (LTSD) between vocal and noisy frames. The implemented system is based in the one presented in (Ramirez et al., 2004), in which an adaptive decision threshold is used in order to get

[1] Universidad Politécnica de Catalunya. http://www.talp.upc.es

the best performance for each signal to noise ratio.
For the calculation of pitch curves a tool developed at AhoLab Signal Processing Group has been used (Luengo et al., 2007). This tool uses dynamic programming with cesptral coefficients in order to estimate the pitch curve.

4. Difficulties encountered

4.1 Scarcity of bilingual speakers

Collecting a database in Basque is not easy, as many people in the Basque Country do not speak Basque or they do not speak it fluently. Table 1 presents the number of Basque speakers in the Basque Autonomous Community in 2001, distributed according to age[2]. In this table active as well as passive bilingual speakers have been considered, i.e. it includes data about those speakers whose primary language at home is Basque and about those whose primary language at home is not Basque. The language competence among the latter is not good in all cases, as they include people who speak Basque with difficulty or do not speak it at all, although they understand or read it well.

Age range	Total	Percentage
16-24	170 453	23.1%
25-34	171 608	23.3%
35-49	175 522	23.8%
50-64	104 055	14.1%
>=65	115 442	15.7%
TOTAL	737 080	100.0%

Table 1: Distribution of active and passive bilingual speakers according to age in the Basque Autonomous Community in 2001.

The knowledge and use of Basque by the inhabitants of the Basque Autonomous Community vary according to the age range. Table 2 shows the percentage of monolingual and bilingual speakers for each age range. The proportion of Basque speakers is higher among young people.
Furthermore, the fact that our bilingual speaker verification database was recorded as an extension of another biometric database that was part of another project has its own drawbacks. The main specifications of the biometric database, such as the number of volunteers, their age distribution and delivery dates had to be respected. As the Basque recordings were not part of the main project, the specifications set for the biometric database did not take into account the special requirements needed to match the goals of the bilingual speaker verification database. It was a priority to fulfil the specifications of the biometric database, even if this meant to deviate from the optimal specifications for the bilingual database. For example, it was not possible to reject a volunteer just because she or he did not speak Basque, as this would had made the recruiting more difficult and extended the delivery dates of the whole database. This is the reason why, although all 55 volunteers recruited were recorded in Spanish , only 30 of them were recorded in Basque, as the remaining ones were not bilingual or fluent in this language.

Age range	Monolingual	Bilingual
16-24	31.4%	68.6%
25-34	50.5%	49.5%
35-49	63.3%	36.7%
50-64	72.5%	27.5%
>=65	67.4%	32.6%

Table 2: Percentage of monolingual and bilingual speakers according to their age range in the Basque Autonomous Community in 2001.

4.2 Deviation from age distribution

In a speaker verification database, the population should be well represented. It is important that the database includes examples representative of all the potential users of the system. This is the reason why this kind of databases are usually balanced in age ranges and gender. To achieve this balance in the recording of the database, a target distribution of speakers is proposed following the expected distribution of potential users, and the selection of contributors to the database is made according to it. Table 3 shows the goal distribution of recordings by age range, and the real distributions achieved among the recorded Spanish and Basque speakers.

Age range	Goal	Spanish	Basque
18 to 25 years	30%	32.7%	33.3%
25 to 35 years	20%	40.0%	53.3%
35 to 45 years	20%	12.7%	10.0%
45 to 55 years	20%	7.3%	3.3%
More than 55 years	10%	7.3%	0.0%

Table 3: Distribution of speaker's age range in the bilingual database.

The recruitment of the speakers was mainly done among the students and staff of the Faculty of Engineering of the University of the Basque Country. The average age in this group is quite low, as reflected in the deviation from the goal distribution that is observed in the 25 to 35 and 45 and up year ranges both for Spanish and Basque. Furthermore, it is greatly difficult to recruit people older than 35 years for a Basque/ Spanish bilingual database, because most of them are monolingual, as shown in Table 2. That is why there are so few Basque speakers in the database in higher age ranges.
The deviation of Basque speakers' distribution from the target values is higher than that achieved for Spanish speakers. Once more, the main reason for this is that during recruitment it was a priority to keep the age distribution for the main biometrical database, in which

[2] Source EAS (Language Indicator System of the Basque Country)
http://www1.euskadi.net/euskara_adierazleak/zerrenda.apl?hizk=i&gaia=25&sel=64

Spanish recordings were included. But then, when non-bilingual people were dropped, the new age distribution for Basque speakers did not match the objective.

The balance in gender was easier to achieve. In table 4 the goal distribution sought in gender and the real distributions both for the Spanish and Basque parts are shown. These real distributions do not differ significantly between Spanish and Basque.

Gender	Goal	Spanish	Basque
Male	50%	47.3%	43.3%
Female	50%	52.7%	56.7%

Table 4: Distribution of speaker's gender in the bilingual database.

5. Conclusions

Taking into account that Basque is a minority language the development of new spoken resources for this language is difficult and the funding for them is usually scarce. Under these circumstances, the acquisition process of a database in a majority language represents an opportunity that can be seized to build another database in the minority language. Using this strategy, a new database for bilingual speaker verification in Spanish and Basque has been created. Its features are not ideal, because the acquisition process has not been designed explicitly for it, but nonetheless it is a new and useful spoken resource. Currently it is being used in the building of a bilingual speaker verification system with success.

6. Acknowledgements

This work has been partially founded by Basque Government under grant IE06-185 (ANHITZ project, http://www.anhitz.com/) and by the University of the Basque Country and EJIE S.A. under grant EJIE07/02 (MULTILOK project).

The authors would also like to thank all the contributors that took part in the acquisition of the biometric database.

7. References

Auckenthaler, R., Carey, M.J., Mason, J.S.D. (2001). Language dependency in text-independent speaker verification. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* vol. 1, Salt Lake City, UT, USA, pp. 441--444.

Bimbot, F., Bonastre, J.F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska-Delacrétaz,D., Reynolds, D.A. (2004) A Tutorial on Text-Independent Speaker Verification. *EURASIP Journal on Applied Signal Processing* vol. 4, pp. 430--451.

Campbell, J.P. (1997) Speaker Recognition: A tutorial. *In Proceedings of the IEEE*, 85, pp. 1437--1462.

Díaz de Ilarraza A., Sarasola K., Gurrutxaga, A., Hernaez, I., Lopez de Gereñu, N (2003). HIZKING21: Integrating language engineering resources and tools into systems with linguistic capabilities. *In Proceedings of the Workshop on NLP of Minority Languages and Small Languages*, Nantes, France.

Faundez-Zanuy, M. Satue-Villar, A. (2006) Speaker Recognition Experiments on a Bilingual Database. *In Proceedings of the 14th European Conference on Signal Processing (EUSIPCO)*, Florence, Italy.

Galbally, J., Fierrez, J., Ortega-Garcia, J. et. al. (2007), BiosecurID: a Multimodal Biometric Database. *In Proceedings of the User-Centric Technologies and Applications Workshop*, Salamanca, Spain, pp. 68-76.

Jain, A.K.; Ross, A.; Pankanti, S. (2006) Biometrics: a tool for information security. , *IEEE Transactions on Information Forensics and Security*, Vol. 1 (2), pp. 125 --143.

Kenny, P.; Dumouchel, P. (2004) Disentangling speaker and channel effects in speaker verification. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* vol. 1, Montreal, Quebec, Canada, pp. 37--40.

Luengo, I., Saratxaga, I., Navas, E., Hernáez, I., Sanchez, J., Sainz, I. (2007) Evaluation Of Pitch Detection Algorithms Under Real Conditions. *In Proceeding of 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, HI, USA, pp. 1057--1060.

Ma, B., Meng, H. (2004) English-Chinese bilingual text-independent speaker verification. *In Proceeding of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)* vol. 5, Montreal, Quebec, Canada, pp. 293--296.

Naik, J.M. (1990) Speaker verification: a tutorial. *IEEE Communications Magazine*, vol. 28(1), pp. 42--48.

Nordstrom, T., Melin, H., Lindberg, J. A Comparative Study of Speaker Verification Systems using the Polycost Database. *In Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)*. Sydney, Australia, pp. 1359--1362.

Ramirez, J., Segura, J., Benitez, C., de la Torre, A., Rubio, A. (2004) Efficient Voice Activity Detection Algorithms Using Long Term Speech Information, *Speech Communication* 42, pp. 271--287.