

# Text independent speaker identification in multilingual environments

Iker Luengo, Eva Navas, Iñaki Sainz, Ibon Saratzaga, Jon Sanchez, Igor Odriozola, Inma Hernaez

University of the Basque Country  
Alda. Urquijo s/n 48013 Bilbao (SPAIN)  
E-mail: {ikerl, eva, inaki, ibon, ion, igor, inma}@aholab.ehu.es

## Abstract

Speaker identification and verification systems have a poor performance when model training is done in one language while the testing is done in another. This situation is not unusual in multilingual environments, where people should be able to access the system in any language he or she prefers in each moment, without noticing a performance drop. In this work we study the possibility of using features derived from prosodic parameters in order to reinforce the language robustness of these systems. First the features' properties in terms of language and session variability are studied, predicting an increase in the language robustness when frame-wise intonation and energy values are combined with traditional MFCC features. The experimental results confirm that these features provide an improvement in the speaker recognition rates under language-mismatch conditions. The whole study is carried out in the Basque Country, a bilingual region in which Basque and Spanish languages co-exist.

## 1. Introduction

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 in another one. Works like those made by Faundez and Satue-Villar (2006) and Durou (1999) show that there is an accuracy decrease in these language-mismatched conditions, but they give no further insight on how to alleviate the problem. Other works like the ones by Akbacak and Hansen (2007) and by Ma and Meng (2004) give some sort of solution to the problem, but they involve knowing in advance the possible languages that will be used, and training a speaker model for each of them, which is not always possible.

In this work we try to find a feature-level solution, i.e. finding a robust parameterization that helps maintaining the recognition accuracy in language-mismatched conditions. Being a feature-level solution, it should be completely generalizable to any language not seen during the training. It addresses this problem in a bilingual region, the Basque Country, in which two official languages co-exist: Basque and Spanish. Both languages have very little in common, as Basque is not an Indo-European language like Spanish, which is a Romance. In fact, Basque is considered a language isolate, i.e. without any relationship with any other living or extinct language. This real-life situation can be seen as the worst-case scenario for a speech processing system, be it a speech recognition or a speaker recognition system, as we are dealing with larger differences than only dialectal ones.

## 2. Definition of the problem

### 2.1 Speaker Identification in Language-mismatched conditions

The most popular approach for speaker recognition systems makes use of Gaussian mixture models (GMM) (Paalanen et al, 2006) of short-term spectral features, like MFCC or LPCC (Young, 1995; Reynolds & Rose, 1995). These spectral features characterize the vocal tract filter at the moment of articulation, effectively capturing not only the characteristic vocal tract of each speaker (thus allowing his/her identification) but also the characteristic vocal tract of each phoneme. This means that MFCC and LPCC features capture information about the phonetic content of the utterance too.

In a text-independent speaker recognition system problems arise when, in a multilingual environment, the model is trained in a language but the testing is made in another. Usually the phonetic content of both languages is not the same, so that the score of the test utterances will be misleading, increasing the error rate of the system.

### 2.2 Proposed solution

A straightforward way to reduce the disagreement between the test utterance and the model is to perform the training with recordings from both languages. In this way, there is a chance for the characteristics of all phonemes to be learnt. This solution is adopted by Ma and Meng (2004). Another way would be having a different speaker model for each language, and using a language detector to decide which one to use, as proposed by Akbacak and Hansen (2007). But these approaches will not be generalizable to a different

language from those used during training. A more language-independent solution is desirable.

In the last few years new high level features have been proposed and successfully applied for speaker recognition problems in a monolingual working frame (Reynolds et al, 2003). Among them, prosodic features, which are related to intonation, energy and speech rate, seem a good option (Dehak, Dumouchel & Kenny, 2007) since they can be easily estimated by automatic signal processing algorithms and can be extracted even from very short utterances. Like the spectral parameterization, these prosodic features contain information of both the speaker and the spoken language. In the case of a multi-language system, the advantages of using prosodic features may arise if the inter-speaker variability of these features is larger than the inter-language variability. In this case it is reasonable to say that prosody is less language-dependent than speaker-dependent. Section 4 describes the measurement of the speaker and language separability of MFCC and prosodic features in order to see if they can be successfully applied to reduce the error rate in our bilingual case.

### 3. Database description

A new Basque-Spanish bilingual speech database was used for the experiments. The database contains recordings of 22 bilingual speakers (11 male and 11 female) in a silent environment. Plantronics DSP-400 microphone was used for this purpose, and audio signals were sampled at 44.1 kHz and 16 bit per sample. Each speaker was recorded in four different sessions, with unequal time spacing, in order to adequately capture the speech variability over time. This bilingual Basque/Spanish database was acquired together with a multimodal biometric database (Galbally et al., 2007) and the calendar designed for the acquisition of the latter was also used for the new database. 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.

Each speaker recorded 7 numeric sequences formed by 8 digits that the speaker read as she or he preferred. All numeric sequences recorded in each session are common for Spanish and Basque.

### 4. Study of the language dependency for the features

In the present work two kinds of features have been used: spectral and prosodic. Traditional MFCC parameterization was used as representative of spectral information. 18 MFCC features augmented with first and second derivatives were calculated every 10 ms, and mean and variance normalization (MVN) was applied to reduce channel effects.

Prosodic features were also considered, as they may improve the performance of the system. Used prosodic features consist of intonation and absolute energy extracted every 10 ms, together with their first and second derivatives. MVN was also applied in order to reduce the inter-session variability of these features. As no pitch information is available in unvoiced frames, only voiced ones are used. This approach makes possible to append prosodic values to the already calculated MFCC vectors, effectively including prosody into the baseline GMM system.

A proper speech parameterization for speaker recognition should have large inter-speaker variability (in order to be able to discriminate among the speakers) and low intra-speaker variability, both inter-language and inter-session (so that the feature distribution does not change much between training and testing conditions). In order to verify the suitability of the proposed features, their variability was estimated using the Kullback-Leibler divergence (Kullback & Leibler, 1951) as a distance measure between distributions.

For inter-speaker variability, the K-L divergence was computed between every possible pair of speakers. The mean value among all the pairs was used as a measure of inter-speaker variability, as it should be representative of the mean divergence between any two random speakers. This estimation was carried out separately for Basque and Spanish.

Similarly, the inter-session variability for each speaker was estimated as the mean K-L divergence between all possible pairs of available sessions for that speaker. The final inter-session variability was estimated as the mean variability among all speakers. This estimation was also carried out separately for Basque and Spanish.

Last, the K-L divergence between the features for Basque and Spanish recordings was calculated for every speaker. Again, the mean value among all speakers was used as a global measure of inter-language variability.

The inter-speaker variability to inter-language variability ratio can be used as a measure to compare two feature sets in terms of language-robustness. In a similar way, the inter-speaker variability to inter-session variability ratio can be a measure of the session-robustness. It is desirable for these two ratios to have a value as large as possible.

Results of all these measures are summarized in Table 1 for plain MFCC and the proposed parameterization. As expected, adding the new prosodic features to the already calculated MFCC vectors increases all variabilities, as new feature dimensions can only increase the distance between two distributions. But while the inter-speaker variability is increased about a 30%, the inter-language variability is increased only a 12%. As reflected by the inter-speaker to inter-language ratio, this results in an increase around 15% in the language-robustness.

		<i>MFCC</i>	<i>MFCC+P</i>	<i>Gain</i>
<b>spk</b>	<b>S</b>	6.34	8.25	30%
	<b>B</b>	6.82	8.77	29%
<b>ses</b>	<b>S</b>	3.62	4.81	33%
	<b>B</b>	3.52	4.64	32%
<b>lang</b>	-	4.09	4.61	12%
<b>spk/lang</b>	<b>S</b>	1.55	1.79	15%
	<b>B</b>	1.67	1.90	14%
<b>spk/ses</b>	<b>S</b>	1.75	1.72	-2%
	<b>B</b>	1.94	1.89	-3%

Table 1: Speaker, session and language KL divergence values for plain MFCC and prosody-augmented (MFCC+P) features, for Spanish (S) and Basque (B) recordings.

The drawback is that intonation and energy measures have a great inter-session variability, so at least part of the gain obtained in language robustness will be lost in session sensitivity. The overall inter-speaker to inter-session variability ratio is reduced around a 2-3%. This means that when language-matched tests are performed (i.e. no inter-language variability occurs), results with the new features are expected to be a little worse than with plain MFCC features.

## 5. Conditions of the experiments

GMM trained with EM algorithm (Duda, Hart & Stork, 2001) were used for both plain MFCC and MFCC vectors enhanced with short-term prosodic features. Plain MFCC models were trained using both voiced and unvoiced frames, while MFCC+short-term-prosody models used only the voiced ones. For comparison purposes, MFCC models with only voiced frames were also evaluated.

Speech material was first downsampled to 8 kHz. A Voice Activity Detector (VAD) based on Long-Term Spectral Deviation (Ramirez et al., 2004) was applied in order to discard silences prior to feature estimation. All 4 available sessions were used for the experiments, in a Leave-One-Out procedure. Each speaker model was trained using two complete sessions (approximately 45 seconds of speech), while a third session was reserved for development, in order to estimate the best values for the meta-parameters of the models (i.e., mixture number). The test sequences were extracted from the last session. In the next iteration all sessions changed their roles. Finally, the mean accuracy among all iterations was calculated as final result. When double-language training was performed, one training session from Spanish and another one from Basque were used, giving a total of two training sessions on the whole. This allows a direct comparison among the systems, as all of them were trained with approximately the same amount of speech.

## 6. Experimental results

As a reference, speaker identification rates of plain MFCC-GMM models trained and tested on the same language are shown in Table 2 (S=Spanish, B=Basque). System’s performance dropped with more than 64 mixtures, as the models overestimated due to the little training data available. Table 3 shows the recognition rates of the 64 mixture models under cross-language conditions. The accuracy drops dramatically when the training and testing language are not the same. But if double-language training is performed, the results are very close to those obtained in the same-language condition.

<i># mix</i>	<i>S-train; S-test</i>	<i>B-train, B-test</i>
<b>2</b>	81.12	79.25
<b>4</b>	89.29	87.93
<b>8</b>	94.05	92.35
<b>16</b>	96.60	95.24
<b>32</b>	97.62	95.41
<b>64</b>	98.34	97.29

Table 2: Identification rates for different number of mixtures when training and testing is performed on the same language and spectral information is used.

<i>S-B</i>	<i>B-S</i>	<i>SB-S</i>	<i>SB-B</i>
63.55	67.34	96.77	95.58

Table 3: Identification rates for the 64 mixture models under cross-language condition. SB means bilingual training.

As stated before, this solution is not generalizable to languages not seen during training, and the use of features robust to language change is preferred. As proposed, we added short-term prosodic features to the MFCC vectors, and checked if they allowed for a better generalization of the models. As pitch values are not defined for unvoiced frames, only voiced frames are considered. Table 4 shows the results with this parameterization. For comparison, results with a plain MFCC system using only voiced frames are also shown. Comparing the results for 64 mixture models in Table 2 and Table 3 with the results in Table 4 for plain MFCC it can be concluded that discarding unvoiced frames has little effect on the results for MFCC parameterization.

When short-term prosodic values are added, the accuracy for same-language condition decreases slightly, as predicted by the feature variability measures in section 4. But cross-language recognition rates improve for monolingual models, as the inter-language

robustness gain is greater than the inter-session robustness loss.

	S-S	B-B	S-B	B-S	SB-S	SB-B
<b>MFCC</b>	97.6	96.8	62.6	67.0	96.6	95.6
<b>MFCC+P</b>	97.1	96.3	71.0	73.0	96.1	94.4

Table 4: Identification rates for the 64 mixture models for voiced frames parameterized with MFCC + short-term prosody.

With multi-language trained models the accuracy is also slightly lower with added prosodic features. These models have already gained language robustness due to the multi-language training. Nevertheless, this robustness is valid only for the two considered languages, namely Spanish and Basque, and the accuracy would drop again if another language is used.

## 7. Conclusions

In this work we have studied the benefits of appending short-term intonation and energy information to traditional MFCC features in order to obtain a more language-independent parameterization for speaker recognition. As a first step, the speaker, session and language variability of these features were estimated. These measurements predicted an improvement of the recognition accuracy in language-mismatch conditions. The results of the speaker recognition tests confirmed this prediction, showing a significant increase of the recognition rates under language-mismatch conditions.

Experimental results also show a slight performance loss when training and testing is done in the same language, due to the great inter-session variability of the prosodic features. Nevertheless, this loss can be acceptable when the system will be used in a multilingual environment and a multilingual training can not be performed, or the languages that the speaker will be using are not known in advance.

Although these short-term prosodic features enhance the language robustness in speaker recognition systems, results are still far from being truly language independent. New features or new language normalization techniques are needed in order to build a system with similar accuracy in language-matched and language-mismatched tests.

## 8. Acknowledgements

This work has been partially funded by Basque Government under grant IE06-185 (ANHITZ project, <http://www.anhitz.com/>) and by the University of the Basque Country and EJIIE S.A. under grant EJIIE07/02 (MULTILOK project).

## 9. References

- Akbacak, M., Hansen, J. H. L. (2007) Language normalization for bilingual speaker recognition systems. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, Honolulu, Hawai'i, pp. 257--260.
- Dehak, N., Dumouchel, P., Kenny, P. (2007) Modeling Prosodic Features With Joint Factor Analysis for Speaker Verification. *IEEE Transactions On Audio, Speech, And Language Processing*, Vol. 15, N. 7, pp. 2095--2103.
- Duda, R. O., Hart, P. E., Stork, D. G. (2001) *Pattern Classification*. Wiley, John and Sons.
- Durou, D. (1999) Multilingual text-independent speaker identification. In *Proceedings of Multi-lingual Interoperability in Speech Technology (MIST)*, Leusden, The Netherlands, pp. 115--118.
- Faundez, M., Satue-Villar, A. (2006) Speaker recognition experiments on a bilingual database. In *Proceedings of IV Jornadas en Tecnologias del Habla (AJTH)*, Zaragoza, Spain, pp. 261--264.
- Galbally, J., Fierrez, J., Ortega-Garcia, J., Freire, M. R., Alonso-Fernandez, F., Siguenza, J.A., Garrido-Salas, J., Anguiano-Rey, E., Gonzalez-de-Rivera, G., Ribalda, R., Faundez-Zanuy, M., Ortega, J.A., Cardeñoso-Payo, V., Vilorio, A., Vivaracho, C. E., Moro, Q. I., Igarza, J.J. Sanchez, J., Hernaez I., Orrite-Uruñuela, C. (2007) BiosecuRID: a Multimodal Biometric Database. In *Proceedings of MADRINET Workshop*, Madrid, Spain, pp. 68--76.
- Kullback, S., Leibler, R. A. (1951) On information and sufficiency. *Annals of Mathematical Statistics*, Vol. 22, N. 1, pp. 79--86.
- Ma, B., Meng, H. (2004) English-Chinese bilingual text-independent speaker verification. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP'04)*, vol. 5, Montreal, Canada, pp. 293--296.
- Paalanen, P., Kamarainen, J. K., Ilonen, J., Kälviäinen, H. (2006) Feature representation and discrimination based on Gaussian mixture model probability densities – Practices and algorithms. *Pattern Recognition*, Vol. 39, N. 7, pp. 1346--1358.
- Ramirez, J., Segura, J. C., Benitez, C., de la Torre, A., Rubio, A. (2004) Efficient Voice Activity Detection Algorithms Using Long Term Speech Information. *Speech Communication*, vol. 42, pp. 271--287.
- Reynolds, D. A. Rose, R. C. (1995) Robust Text Independent Speaker Identification using Gaussian Mixture Speaker Models. *IEEE transactions on Speech and Audio Processing*, vol. 3, pp. 72--83.
- Reynolds, D. A., Campbell, J. P., Dunn, R. B., Gleason, T., Jones, D., Quatieri, T. F., Carl, Q., Sturm, D., Torres-Carrasquillo, P. (2003) Beyond Cepstra: Exploiting High-Level Information in Speaker Recognition. In *Proceedings of Workshop on Multimodal User Authentication*, Santa Barbara, CA, USA, pp. 223--229.
- Young, S. (1995) Large vocabulary speech recognition: A review. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, Snowbird, Utah, pp. 3--28.