# Comparative Study of Several Phonotactic-Based Approaches to Spanish-Basque Language Identification*

Víctor G. Guijarrubia and M. Inés Torres

Departamento de Electricidad y Electrónica
Universidad del País Vasco, Apartado 644, 48080 Bilbao, Spain
{vgga,manes}@we.lc.ehu.es

**Abstract.** This paper presents a series of language identification (LID) experiments for Spanish and Basque. Spanish and Basque are both official languages in the Basque Country, a region located in northern Spain. We focused our research on studying several phonotactic-based methodologies, comparing both the performance of phonotactic models trained from text and audio samples and the use of phone and phone-sequences as decoding units. The results show that whereas the use of audio-based phonotactic models performs better than the text ones, when using task-specific information it is also possible to achieve great accuracies. The use of phone sequences as decoding units appears to be useful when constraining the phone decoders to those sequences.

**Keywords:** language identification, phone decoding, pprlm.

## 1 Introduction

Language identification (LID) is a classical pattern recognition problem that is strongly tied to multilingual speech recognition and dialogue systems.

The typical LID system consists of one or more speech tokenizers that convert the input utterance into a sequence of tokens that can be evaluated by some stochastic models to produce some scores. These scores are then used to take a decision. A widely used approach is the parallel phone recognition followed by n-gram language modelling (PPRLM) technique [1]. In this case, some phoneme decoders are used to tokenize the input sequence, which is then analyse by phonotactic models to generate some scores and predict the spoken language. Another approach is to use Gaussian mixture models as tokenizers [2].

Some scientific evaluations include the identification of a language among a set of languages [3]. However, for multilingual communities high LID performances are required, but only for the involved languages, typically two or three. This was the goal of our work. This paper focus on robust identification of Basque

* This work was partially supported by the Spanish CICYT project TIN2005-08660-C04-03 and by the University of the Basque Country under grant GIU07/57.

J. Ruiz-Shulcloper and W.G. Kropatsch (Eds.): CIARP 2008, LNCS 5197, pp. 128–135, 2008.

and Spanish languages. Basque is a minority language, but it is the joint official language, along with Spanish, for the 2.5 million inhabitants of the Basque Country (Spain). Basque is a preindoeuropean language of unknown origin.

Basque and Spanish present many differences at lexical, morphological and syntactical level. However both languages share important acoustic-phonetic characteristics and their acoustic discrimination turns out to be a not so easy task. Basque and Spanish languages share the same vowel triangle that includes only five vowels. The set of Basque phones does not differ much from the Spanish ones. Nevertheless, Basque includes larger sets of fricative and affricate sounds.

In this work we want to compare the performance of phonotactic models trained from text samples and from audio samples. For the PPRLM technique, the phonotactic models are trained from the output of some acoustic decoders. The reason to try text-based phonotactic models is to evaluate if the use of text samples is good enough, so it is possible to use them to overcome the problem of acoustic resources availability present especially in minority languages like Basque. We are also interested in testing the use of phone sequences as basic units in phonotactic-based LID systems. The motivation for this is to take advantage of sequences of sounds that appear frequently in each language. A explicit modelling of these sequences could help phonotactic-based discrimination, adding a new level between phones and words. To obtain those sequences, we propose a simple technique based on *n-gram* statistics.

## 2 Language Identification Methodologies

We describe here the basic LID techniques used in all the experiments and how to get and use sequences of phones as basic units.

### 2.1 Basic Language Identification Techniques

In order to perform the proposed language identification task, some phone decoding methods were implemented. These techniques rely on acoustic phonetic decoders, which find the best sequence of decoding units depending on the input speech signal. In our case, these decoders are based on the Viterbi algorithm, which, given an input, finds the most likely path through a probabilistic network. When applied to an acoustic phonetic decoder, this network consists of a combination of all the acoustic models, them usually being Hidden Markov Models (HMMs) associated to a previously defined set of phonetic units of the language. In this sense, given a set of acoustic models $\Lambda^l$ associated to a language $l$ and an input sequence of acoustic observations $O = o_1 \dots o_T$, a Viterbi decoder finds the best sequence of states $Q^l = q_1 \dots q_T$ through the network of models. This can be expressed in a mathematical manner as follows:

$$Q^l = \underset{q_1 \dots q_T}{\arg\max}\, P(q_1 \dots q_T, o_1 \dots o_T | \Lambda^l) \tag{1}$$

The path $Q^l$ determines a sequence of decoding units $\hat{X}^l = x_1^l \dots x_N^l$, based on the previously defined set of HMMs associated to language $l$. The following paragraphs describe one by one the different techniques that were explored.

**Parallel Phone Recognition and Language Modelling (PPRLM)**

PPRLM is one of the most widely used technique in language identification. As reported before, some phoneme decoders (in our experiments, Basque and Spanish decoders) are used to tokenize the input sequence, which is then analyse by phonotactic models to generate some scores and predict the spoken language.

In this case, audio samples are used to train the phonotactic models answering the question, "What is language $i$ according to the acoustic models of language $j$?". Mathematically, it can be expressed as

$$L = \arg\max_l P(l) = \arg\max_l \sum_{l'} \log P(\hat{X}^{l'} | Ph_{l'}^l) \tag{2}$$

where $\hat{X}^l$ is the sequence of units estimated by the acoustic decoder for language $l$ and $Ph_{l'}^l$ represents the phonotactic model for language $l$ according to the acoustic model of language $l'$.

The benefits of using this technique is that you can develop a LID system for the languages you want, given some acoustic models, since no knowledge is required. However, it requires audio samples to train the phonotactic models, which is not always easy to find, and the recording conditions or type of information of the samples could not be suitable for the desired application.

**Phone decoder scored by a phonotactic model (PD+PhM).** Another option is to train the phonotactic models for text samples rather than speech. In this case, for every language being studied, an unconstrained acoustic decoder is applied, resulting in a sequence of decoding units for each language. A language-dependent phonotactic model is then employed to assign a score to each of the sequences for that language. The highest score identify the language according to

$$L = \arg\max_l P(l) = \arg\max_l P(\hat{X}^l | Ph^l) \tag{3}$$

where $\hat{X}^l$ is the sequence of units estimated by the acoustic decoder and $Ph^l$ represents the phonotactic model for language $l$.

**Phone decoder constrained by a phonotactic model (PDPhM).** Also known as PPR in the literature [1], this method performs a phone decoding for each language being studied, but constrained by a phonotactic model. Thus, in this case, the phonotactic model is used during the decoding process, whereas in the previous techniques was applied after the decoding.

This way, the decoder is similar to a speech recognition system. In this case, our goal is to find a sequence of phonetic units instead of a sequence of uttered

words. In this context, the best sequence of decoding units $X^l$ that fits the input sequence of acoustic observations $O$ is found applying the Bayes' rule

$$P(X^l|O) = P(O|X^l)P(X^l)/P(O) \tag{4}$$

where $P(O|X^l)$ is the probability of the acoustic sequence for that particular phonetic string; this value is computed using the acoustic models. $P(X^l)$ is the *a priori* probability of the sequence of decoding units, and is computed using a phonotactic model. In the same way, $P(O)$ represents the *a priori* probability of the acoustic sequence. Typically this parameter is not computed, since it has a constant value across all the possible phonetic strings obtained from a given decoding. However, when comparing the output of different recognisers, this probability should also be considered. In this work, we approximated that term using an acoustic normalisation, in a similar way as that presented in [4]. The acoustic likelihood of each of the decoded units is normalised by the likelihood of the best unconstrained phone sequence in that period of time.

Finally, the hypothesised language is assumed to be the one for which

$$L = \arg\max_l P(l) = \arg\max_l P(X^l|O) \tag{5}$$

### 2.2 Phone-Sequences Based LID

As reported before, we wanted to test the use of phone sequences in phonotactic-based LID systems. The fundamental idea of using phone sequences is to take advantage of sequences of sounds that appear frequently in each language, like prefixes or suffixes, hopping that constraining the systems this way we add a new level of knowledge, but without reaching the word level.

This process is summarised in Algorithm 1.

These sequences were used in the phonotactic models. For the PDPhM technique, they are directly employed during decoding. For the other techniques, the decodings are performed using phone-like units and the resulting strings are then re-labelled using the final list of sequences, so they can be analysed by sequences-based phonotactic models.

### 2.3 Bias Removal

Due to differences in the number of phone-like units for each language, the language with the smallest inventory of units will tend to dominate the other languages because of higher phonotactic scores. To remove the bias this introduces, we used a technique suggested in [1]. Given some input utterances from all languages, this method consists of finding a language-dependent bias $b(l)$ as the average of the log-likelihood score for those input utterances. Supposing there are $s$ input utterances per language, then

$$b(l) = \frac{1}{N * s} \sum_{i=1}^{N*s} P^i(l) \tag{6}$$

where $N$ is the number of languages and $P^i(l)$ is the score assigned by the

**Algorithm 1.** Phone_sequences algorithm

**Input:** sampling corpus, maximum length of the sequences (N) and minimum number of occurrences (MIN)

**Output:** sequences that appear simultaneously in the given corpus at least MIN times

1: Get all the $n$-grams ($n = 2, \ldots, N$) and their number of occurrences present in the sampling corpus.
2: Sort the list of $n$-grams in order of decreasing values of $n$, decreasing number of appearances and according to inverse alphabetical order.
3: Label the corpus with the resulting list, i.e. replace all the appearances of a sequence of phones corresponding to a $n$-gram with a single unit obtained joining all the phones forming that $n$-gram.
4: **for all** sequence in the list **do**
5: **if** the number of occurrences of the sequence in lower than MIN in the labelled corpus **then**
6: Delete the sequence from the list.
7: Relabel the samples of that sequence using the remaining sequences.
8: **end if**
9: **end for** ▷ The final list is the output of the algorithm.

LID method to the language $l$ when taking the $i$-th utterance as input. Thus, equations 2, 3 and 5 can be expressed as

$$L = \arg\max_{l} [P(l) - b(l)] \tag{7}$$

It is also assumed that the utterance log-likelihood scores are normalised by the length of the decoded phone sequence (for PPRLM and PD+PhM) and by the length of the utterance (number of frames) plus the length of the decoded phone sequence for the PDPhM technique.

## 3 Task and Corpora

To carry out the identification experiments, some speech databases were required, all of then composed of read speech recorded at 16 `kHz`.

### 3.1 Evaluation Database

The evaluation set consisted of a weather forecast database called `METEUS` [5]. The text data was picked from the Basque weather service *Euskalmet* for Spanish and Basque. A subset of 500 different sentences was selected and recorded by 36 speakers for each language. The 500 sentences were divided into blocks of 50 sentences each and every speaker uttered the sentences corresponding to one of these blocks. A total of 1800 utterances were recorded for each language.

### 3.2 Training Speech Data

The LID systems used in this work are based on phone recognisers. Thus, a basic set of acoustic models representing the phones of each language needs to

be trained. Some acoustic databases containing phonetic transcription are thus required for each language:

- Basque: a phonetically balanced database called EHU-DB16 [6] was used to train the acoustic models. Not only is phonetically balanced in terms of phone occurrences, but also in terms of phone contexts. This database contains 9394 sentences uttered by 26 speakers. It also includes a testing subset of 1056 utterances from 14 speakers.
- Spanish: we resorted to the phonetic corpus of the Albayzin database [7]. Also phonetically balanced, it consists of 4800 sentences uttered by 164 speakerses. There is also a test subset composed of 2000 utterances from 40 speakers.

### 3.3 Phonotactic Resources

The LID systems also need some phonotactic models to represent the combinations of sounds that occur in each language.

The PPRLM technique needs audio resources to train the phonotactic models. The testing subsets from the EHU-DB16 and Albayzin were used for that. The acoustic models were used to decode these subsets and the required phonotactic models were training from the resulting strings.

For the other techniques, text resources are needed. In this case, two options were explored: generic and task-specific phonotactic model. For the generic models, the phonetic transcriptions of the EHU-DB16 and Albayzin databases were used. Being them phonetically balanced, they should represent with high fidelity the real phonotactic of each language, both in terms of number of occurrences and phonetic contexts. For the task-specific model, a set of 14615 task-related sentences for each language were used. None of the sentences of the evaluation database were include in these text resources.

## 4 Experimental Results

### 4.1 Experimental Conditions

Within the frame of the experiments that were carried out, the databases were parametrised into 12 Mel-frequency cepstral coefficients with delta and acceleration coefficients, energy and delta-energy. The length of the analysis window was 25 ms and the window shift, 10 ms.

Each phone-like unit was modelled by a typical left-to-right non-skipping self-loop three-state HMM, with 32 Gaussian mixtures per state. A total of 35 context-independent phones were used for Basque and 24 for Spanish. For the PDPhM-sequences method, the acoustic models were build concatenating the models of their constituent phones.

For the above-mentioned LID techniques, a phonotactic model is also required to score the recognised phone sequence. In these experiments, only trigram models were evaluated .

### 4.2 Results of the Experiments

First of all, for every language and technique we needed to calculate the language-dependent bias. For the PD+PhM and PDPhM techniques, the test subsets of the Albayzin and EHU-DB16 were used. In order to use the same amount of utterances for each language, only the first 1056 utterances of the Albayzin test were used. For the PPRLM technique, since the testing subsets were used to train the phonotactic models, the training subsets were used to calculate the bias. In this case, only the first 4800 utterances of the EHU-DB16 training subsets were employed.

In order to carry out the experiments, a complete utterance was presented to the LID system, implementing the various approaches described in Section 2. The results, in terms of LID accuracy, are summarised in Table 1.

**Table 1.** LID accuracies, i.e. the number of correctly classified utterances, for generic and task-specific phonotactic models, phone-based and sequences-based phonotactic models, and according to the techniques described in Section 2

| | Generic PhM | | Specific PhM | |
|---|---|---|---|---|
| | Phone | Seq. | Phone | Seq. |
| PD+PhM | 83.4 | 71.6 | 97.5 | 83.0 |
| PDPhM | 81.1 | 83.6 | 98.6 | 99.9 |
| PPRLM | 91.1 | 86.2 | | |

As mentioned above, the aims of the present work was to asses the performance of phone-sequences based systems versus those systems that rely on phones only and to compare the performance of text-based phonotactic models against audio-based models. Starting with the later, it is clear from the results that using generic phonotactic models, the PPRLM outperforms the other techniques. Considering that during the decoding process there are some mistakes and the resulting string is not perfect, training the phonotactic models from the output of phone decoders helps the PPRLM system to perform well. The PD+PhM technique is more susceptible to the mistakes introduces by the decoders and that results in more errors. From the results we can see that using phones as units the PDPhM does not perform as well as the other techniques. We can conclude then that the evaluation database is not phonetically balanced, so we are constraining the system to something is not appropriate and this introduces more mistakes. This could also be affecting to the other techniques. The use of task-specific phonotactic models results as expected in a great improvement for both PD+PhM and PDPhM. In this case, the PDPhM performs better than PD+PhM

Comparing the performance of phone-sequences and phone based systems, we can see that only the PDPhM technique improves under that situation. The uttered language benefits from the acoustic scores due to more reliable paths assigned by the phonotactic model. For the PD+PhM technique, on the other hand, does not get any benefit from the sequences. The mistakes produced by

the decoders could be affecting the performance, since the phone-sequences are based on real samples. For the PPRLM method, the use of sequences does not perform well either. In this case, the reason could also be the unbalance between the data used to train the phonotactic models and the evaluation database, since we are modelling a phonetically balanced database and the evaluation database does not look to follow that balance.

## 5 Concluding Remarks

In this paper we have studied the performance of several phonotactic-based LID techniques under different conditions, like text and speech based pronotactics or sequences based models. The speech-based phonotactic models let you develop a LID system for the languages you want, but acoustic resources are scarce for minority languages and this can be overcome using text-based models.

The results show that when using generic phonotactic models, the PPRLM performs much better that the text-based techniques. However, when using task-specific phonotactic models, the text-based approaches have a great performance, of almost 100%. The sequences-based models appear to be useful when used during decoding. This lead to a robust system with a 99.9% of accuracy when using a task-specific phonotactic model.

## References

1. Zissman, M.A.: Comparison of four approaches to automatic language identification of telephone speech. IEEE Transactions on Acoustics Speech and Audio Processing 4(1), 31–44 (1996)
2. Torres-Carrasquillo, P.A., Reynolds, D.A., Deller, J.R.: Language identification using gaussian mixture model tokenization. In: ICASSP, Orlando, vol. 1, pp. 757–760 (2002)
3. Martin, A.F., Le, A.N.: The current state of language recognition: Nist 2005 evaluation results. In: IEEE Odyssey 2006, the Speaker and Language Recognition Workshop, San Juan, Puerto Rico, pp. 1–6 (2006)
4. Young, S.R.: Detecting misrecognitions and out-of-vocabulary words. In: ICASSP, Adelaide, Australia, vol. 2, pp. 21–24 (1994)
5. Pérez, A., Torres, I., Casacuberta, F., Guijarrubia, V.: A Spanish-Basque weather forecast corpus for probabilistic speech translation. In: 5th SALTMIL Workshop on Minority Languages, Genoa, Italy, pp. 99–101 (2006)
6. Guijarrubia, V., Torres, I., Rodriguez, L.J.: Evaluation of a spoken phonetic database in basque language. In: LREC, Lisbon, vol. 6, pp. 2127–2130 (2004)
7. Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J.B., Nadeu, C.: Albayzin speech database: Design of the phonetic corpus. In: EUROSPEECH, Lisbon, vol. 1, pp. 175–178 (1993)