Phone-Segments Based Language Identification for Spanish, Basque and English*

Víctor 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, Basque and English. Spanish and Basque are both official languages in the Basque Country, a region located in northern Spain. We focused our research on some techniques based on phone decoding. We propose the use of phone segments as decoding units instead of just phones. We describe a simple procedure to obtain a set of phone segments that typically appear in the languages involved. In comparison with similar techniques that do not rely on phone segments, the choice of these segments as decoding units yields a remarkable improvement in terms of LID accuracy: from 93.02% using phones to 98.32% using phone segments, when applied to trilingual read speech.

Keywords: language identification, phone decoding.

1 Introduction

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

It has been addressed in the past using a variety of tactics; for instance, those exploiting prosodic cues [1] as rhythm or intonation. Nevertheless, most of them are based on speech recognition approximations: phone decoding approaches [2,3], which rely on phone sequences; Gaussian mixture models [2,4] treating only the acoustic; or large-vocabulary continuous-speech recognition approaches [5], which operate based on full lexical sequences. A thorough analysis discussing the current state of the LID systems can be consulted here [6].

The typical LID system is based on a phone recognition followed by n-gram language modelling (PRLM) or, most commonly, parallel PRLM (PPRLM) [2]. In these cases, some monolingual phoneme decoders are used to tokenise the input sequence, which is then analysed by phonotactic models to predict the spoken language. Although most of these systems use language-dependent phonemes, there are some recent works dealing with unified phoneme sets [7].

* This work was partially supported by the Spanish CICYT project TIN2005-08660-C04-03 and by the University of the Basque Country under grant 9/UPV 00224.310-15900/2004.

L. Rueda, D. Mery, and J. Kittler (Eds.): CIARP 2007, LNCS 4756, pp. 106–114, 2007.

The ultimate goal of any LID system is to identify the language being used by an unknown speaker. In some evaluations, like those proposed by the National Institute of Standards and Technology (NIST), 12 or 7 languages are included in those LID systems [6]. However, for multilingual communities high performances are required, but only for the involved languages, typically two or three.

The aim of this work is to build a LID system for Spanish, Basque and English. 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 (northern Spain).

The main differences between Spanish and Basque fall on the lexical units and the morphosyntactic structure. From a phonetic point of view, the set of Basque phones does not differ much from the Spanish one. The two languages share the same vowels (only five). Nevertheless, Basque includes larger sets of fricative and affricate sounds. English, on the other hand, is phonetically very different from Spanish and Basque and includes a larger number of vowel and semi-vowel sounds. In addition, the way to get the phonetic transcription is also different. Whereas for Spanish and Basque the phonetic transcription can be generated by means of a simple set of rules, English transcriptions require the use of a dictionary. Thus, we could presume that English could be discriminated from Spanish and Basque using only acoustic features. However, as suggested in [**?**], a Basque-Spanish discrimination would require information about how the phones combine in each language.

In this paper, we propose the use of phone segments as the decoding units of a LID system. The fundamental idea is to take advantage of sequences of sounds that appear frequently in each language, with the purpose of improving the phone decoding rates and in order to better identify the language being used. To obtain those segments, we propose a simple technique based on N-gram statistics.

In this sense, the remainder of the paper is organised as follows: Section 2 presents the procedure applied to obtain the phone segments, Section 3 describes each of the LID methods used in this study, Section 4 centres on the main features of the speech databases used in the experiments, Section 5 presents the results obtained for the different LID approaches, comparing LID accuracy values for both phones and phone segments, and finally Section 6 discusses the conclusions of the present work.

2 Obtaining the Phone Segments

We propose the use of phone segments as decoding units, with the idea of getting a better representation of each language. To obtain those segments, a simple procedure based on N-gram statistics was used. This process is summarised in the following points:

- Given the training corpus, identify and extract all the 2-grams, 3-grams, ..., n-grams available. In our case, we chose $n = 5$, because it takes into account the most common prefixes, suffixes and words appearing in the languages.
- Sort them in order of decreasing values of n (5-grams before 4-grams, 4-grams before 3-grams, ...), decreasing number of appearances and according

to inverse alphabetical order. This final condition appears naturally when sorting the n-grams in decreasing number of appearances using the *sort* GNU/Linux command.
- Get the subset of phone n-grams that, while keeping the original order, satisfies a minimum number of occurrences. The idea is to 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. Some of the phone n-grams might not appear after this process or might not satisfy the minimum number of occurrences, due to the fact that they could be included in previous phone n-grams. The first of those n-grams not satisfying the minimum number of occurrences is then removed. The process of relabelling and search for not valid n-grams is iteratively repeated until getting the final subset.

3 Language Identification Methods

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, usually being them 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 Λ^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 $Q = q_1 \dots q_T$ through the network of models. This can be expressed in a mathematical manner as follows:

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

The path Q determines a sequence of decoding units $X^l = X_1 \dots X_N$, based on the previously defined set of HMMs associated to language l. In this work, we decided to evaluate phones against phone segments as decoding units to assess their impact upon the accuracy of the associated LID system.

The following subsections describe one by one the different techniques that were explored.

3.1 Phone Decoder Scored by a Phonotactic Model (PD+PhM)

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 language of the utterance is selected to be that with the highest score; that is, the language for which

$$L = \arg\max_{l} P(X^l|Ph^l) \tag{2}$$

where Ph^l represents the phonotactic model for language l. Typically, these phonotactic models are modelled using *n-grams*. Thus

$$P(X^l|Ph^l) = \prod_{i=1}^{N} P(X_i^l|X_{i-1}^l, \ldots, X_{i-n+1}^l, Ph^l) \tag{3}$$

A block diagram of the PD+PhM technique is shown in Figure 1. This technique could be considered as a simplification or variation of the commonly used PPRLM technique.

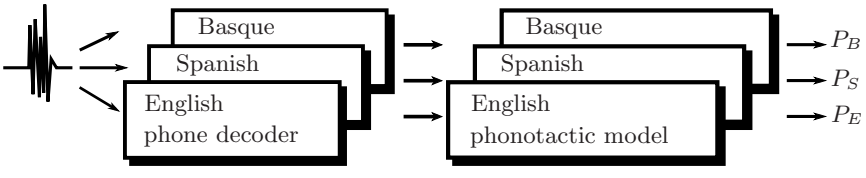

Fig. 1. PD+PhM block diagram. P_B stands for Basque probability, P_S stands for Spanish probability and P_E stands for English probability.

3.2 Phone Decoder Constrained by a Phonotactic Model (PDPhM)

Also known as PPR in the literature [2], this method performs a phone decoding for each language being studied, but constrained by a phonotactic model. That is, in this case, the phonotactic model is used during the decoding process, whereas in the PD+PhM 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 HMMs. $P(X^l)$ 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 lexical 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 (referred as an acoustic confidence measure), in a similar way as that presented in [9]. This technique reported improvements in other

LID applications [10]. 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

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

A block diagram of the PDPhM technique is shown in Figure 2.

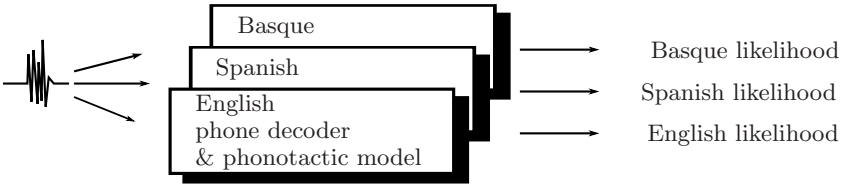

Fig. 2. PDPhM block diagram

4 Speech Corpora

The experiments reported in this paper were performed using several speech databases.

The training of the basic acoustic models for Basque was carried out by means of a phonetically balanced database called EHU-DB16 [11]. This database contains 9394 sentences uttered by 25 speakers and includes around 340000 phones. The resulting models reported phone recognition accuracies of around 74% for this database.

For Spanish, we resorted to the phonetic corpus of the Albayzin database [12], consisting of 4800 sentences uttered by 29 speakers, resulting in around 187000 phones and also being phonetically balanced. The resulting models reported phone recognition accuracies of around 75% for this database.

For English, we chose the *Wall Street Journal 1* database (the SI200 corpus, to be precise). It is composed of more than 30000 sentences uttered by 200 speakers, resulting in more than 66 hours of speech material with around 2 million phones. The resulting models reported phone recognition accuracies of around 58% for this database.

The evaluation set consisted of a weather forecast database recorded initially for Spanish and Basque [13] and later for American English. This database contains 500 different sentences uttered by 36 speakers for every 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. Table 1 summarises the main features of this database. Although there are some spontaneous effects, the data sources are read speech.

Table 1. Main features of the evaluation database

	Spanish	Basque	English
Speakers	36		
Utterances	1800		
Length (hours)	3	3.5	3.4
Average Length of an utterance (sec)	6	7	6.8

It is important to mention that not only do the three languages share the same task and recording conditions, but also two of the languages (Spanish and Basque) share the same speakers. This reduces possible effects benefiting one language from another. Another important aspect to take into account is that silences were not removed from the utterances.

5 Experimental Results

5.1 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. Thus, four acoustic representations were defined. 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 and acoustic representation. The phone sets were based on the phonemes of each language. A total of 35 context-independent phone-like units were used for Basque, 24 for Spanish and 25 for English. This reduced set of 25 units for English is based on the 39 phone set used by the Carnegie Mellon University (CMU) in its pronouncing dictionary. A previous study was carried out to improve the acoustic decoding accuracies over the Timit database; in this sense, and based on the confusion matrices, some units were merged, leading to the definitive set of units being used. The phone recognition accuracies improved from 59.97 to 65.46. For the segments, 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. Moreover, in order to adhere to the phonetic constraints, a *k-testable in the strict sense* (k-TTS) model [14] was used throughout these experiments. The k-TTS are similar to variable-length n-grams, with k and n having approximately the same meaning. Different k values (ranging from $k = 3$ to $k = 5$) were evaluated.

These phonotactic models were trained using several text corpora available at our disposal. For Basque and Spanish, these corpora were phonetically

transcribed based on rules developed by experts, whereas for English the transcription was done using a dictionary; more precisely, the *CMU pronouncing Dictionary* (version 0.6).

5.2 Results of the Experiments

First of all, for every language we needed to obtain a basic set of decoding units consisting in phone segments. For this purpose, the process described in Section 2 was applied. As the training material for each language is different, the minimum number of appearances required to each language was also different. The idea was to get, initially, a similar number of segments for all three languages (around 500). For Spanish and Basque this minimum threshold value was set to 1000 whereas for English it was set to 4000.

Once applied the process described in Section 2, the number of decoding units was 172 for Spanish, 321 for Basque and 221 for English. These units were also used to train the phonotactic models for the segment-based approaches. That is, the phonotactic models of the segment-based approaches are n-grams of phone segments.

In order to carry out Spanish-Basque-English identification experiments, a complete utterance was presented to the LID system, implementing the various approaches described in Section 3.

As mentioned above, one of the aims of the present work was to asses the performance of phone-segment based systems versus those systems that rely on phones only. The results, in terms of LID accuracy, are summarised in Table 2. It is worth pointing out that for both techniques a decoded-string length normalisation was used, since this approximation yielded the best results. Only the PDPhM technique has been applied when using the phone segments (denoted as PDPhM(s) in Table 2). The reasons for this is that the advantage provided by the phone segments is that they help the uttered language while making worse the other languages due to poorer acoustic scores. When using the PD+PhM technique, as the system is not constrained, it is not forced to go through the segments and no real advantage is achieved.

As can be seen, the use of phone segments as decoding units results in a great improvement. Using better phonotactic models, phone segments can yield accuracies of nearly 99%.

One of the differences between the PD+PhM and the PDPhM technique is that the PDPhM includes acoustic scores. Looking at the results for the PD+PhM and PDPhM using phones, we can see that Spanish and Basque benefit form these acoustic scores, whereas English does not. This can be explained by the fact that the Spanish and Basque HMMs are better estimated because they are trained using more reliable phonetic transcriptions. For example, for the PDPhM technique and $k = 4$, the phone recognition rates are around 85% for Spanish and Basque, but only around 60% for English. Note also that whereas the Spanish and Basque transcriptions are completely reliable, the English ones are not. However, for $k = 5$, PD+PhM performs worse than PDPhM. Further investigation should carried out to explain this fact.

Table 2. LID accuracies values for several phonotactic models and according to the techniques described in Section 3

	k	Spanish	Basque	English	Overall
PD+PhM	3	91.71	91.83	80.72	88.09
	4	91.16	94.17	80.72	88.68
	5	92.94	95.17	73.22	87.11
PDPhM	3	99.83	93.72	47.61	80.39
	4	99.89	98.17	63.39	87.15
	5	99.89	98.94	80.22	93.02
PDPhM(s)	3	99.89	99.61	87.83	95.78
	4	99.89	99.67	95.33	98.30
	5	99.89	99.56	95.50	98.32

The use of phone segments improve the results for all the languages. Even if it looks that restricting the decoder is worse for English, when using the phone segments a significant improvement is achieved. For Spanish and Basque the benefits are small, mainly because of the already high accuracies. As commented before, when using the segments, the acoustic scores assigned to the non-uttered languages are much more small, due to they are being forced by the phonotactic model through some predefined paths. However, the uttered language benefits from more reliable paths assigned by the phonotactic model. For example, for $k = 4$ the phone recognition rates in this case are around 95% for Basque and Spanish and around 75% for English. The results clearly demostrate that the phone segments are useful for languages with poorer acoustic modelling. In this work that happened for English, but for other tasks or languages, that could happen for other languages. This also reinforces the idea of exploring a unified phoneme set to overcome similar problems.

6 Concluding Remarks

In this paper we have presented a simple procedure to gather some phone-segments. The use of these phone segments as decoding units resulted in a notable improvement of the associated LID system in terms of accuracy: comparing the results to those obtained using only phones as decoding units, the accuracy increased from 93.02 to 98.32%. The effect of these phone segments is especially significant for English, allowing a remarkable increase in the accuracies. The phone segments help modellize better the language being uttered and worse the others, providing the improvement in the LID accuracies. The phone segments are useful for English in this work, but under different conditions, they could be helpful for others as well.

References

1. Itakahashi, S., Du, L.: Language identification based on speech fundamental frequency. In: EUROSPEECH, Madrid, Spain, vol. 2, pp. 1359–1362 (1995)
2. Zissman, M.A., Singer, E.: Automatic language identification of telephone speech messages using phoneme recognition and n-gram modelling. In: ICASSP, Adelaide, Australia, vol. 1, pp. 305–308 (1994)
3. Navrátil, J., Zühlke, W.: An efficient phonotactic-acoustic system for language identification. In: ICASSP, Seattle, USA, vol. 2, pp. 781–784 (1998)
4. Singer, E., Torres-Carrasquillo, P.A., Gleason, T.P., Campbell, W.M., Reynolds, D.A.: Acoustic, phonetic and discriminative approaches to automatic language identification. In: EUROSPEECH, Geneva, Switzerland, pp. 1349–1352 (2003)
5. Schultz, T., Rogina, I., Waibel, A.: Lvcsr-based language identification. In: ICASSP, Atlanta, USA, pp. 781–784 (1996)
6. Martin, A.F., Le, A.N.: The current state of language recognition: Nist 2005 evaluation results. In: Proceedings of the IEEE Odyssey 2006, the Speaker and Language Recognition Workshop, San Juan, Puerto Rico (2006)
7. Li, H., Ma, B.: A phonotactic language model for spoken language identification. In: ACL 2005, Morristown, NJ, USA, pp. 515–522 (2005)
8. Guijarrubia, V., Torres, I.: Basque-spanish language identification using phonebased methods. In: Proceedings of International Conference of Spoken Language Processing, Pittsburgh, USA, pp. 1780–1783 (2006)
9. Young, S.R.: Detecting misrecognitions and out-of-vocabulary words. In: ICASSP, Adelaide, Australia, vol. 2, pp. 21–24 (1994)
10. Hieronymus, J.L., Kadambe, S.: Spoken Language Identification Using Large Vocabulary Speech Recognition. In: Proceedings of International Conference of Spoken Language Processing, Philadelphia, USA, pp. 1780–1783 (1996)
11. Guijarrubia, V., Torres, I., Rodríguez, L.J.: Evaluation of a Spoken Phonetic Database in Basque Language. In: LREC 2004, Lisbon, vol. 6, pp. 2127–2130 (2004)
12. 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 (1993)
13. 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)
14. Torres, I., Varona, A.: K-TSS Language Model in a Speech Recognition System. Computer Speech and Language 15(2), 127–149 (2001)