Actas V Jornadas TIMM, pp 7-10

Language identification with limited resources

Emilio Sanchis Mayte Giménez Lluís-F. Hurtado

Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València, València, Spain
{esanchis, mgimenez, lhurtado}@dsic.upv.es

Abstract

Language identification is an important issue in many speech applications. We address this problem from the point of view of classification of sequences of phonemes, given the assumption that each language has its own phonotactic characteristics. In order to achieve this classification, we have to decode the speech utterances in terms of phonemes. The set of phonemes must be the same for all the languages, because the goal is to have a comparable representation of the acoustic sequences. We followed two different approaches using the same acoustic model: we decode the audio using trigrams of sequences of phonemes and equiprobable unigrams of phonemes as language model. Then a classification process based on perplexity is performed.

1 Introduction

Language identification (LI) is an important application in multilingual speech environments. This is the case of multilingual dialog systems where the system has to detect the input language in order to choose the corresponding models associated to each language. Given the interest of this field in speech technologies some evaluation campaigns have been proposed, as the Albayzin evaluation in Spain [Rod13]. Some methodologies are used for language identification, some of them directly based on acoustic representation of the signal, and others based on phonetic representations [Pal13]. Our approach consist of a first process of Acoustic-Phonetic Decoding (APD), considering the set of Spanish phoneme models, and a classification process of the sequences of phonemes based on the distance to the different languages. An advantage of this approach is that it can be easily developed when there are not many resources to learn accurate acoustic representation for each language. It is enough to have a set universal phonemes, and a not labeled corpus of each language. We have applied this approach to a multilingual version of the DIHANA corpus, that consist of dialogs for obtaining information about trains in Spain. We present some experiments over English, French and Spanish.

2 Our language identification approach

Our proposal to LI is based on modeling sequences of phonetic units that characterize each language we want to identify. The language identification process of a spoken utterance is divided into two phases:

- Acoustic-Phonetic Decoding. The first phase of the LI process is a phonetic transcription of the spoken utterance which language must be identified. In our proposal, this phase is the same for all languages and, therefore, it should be language independent.

In: L. Alfonso Ureña López, Jose Antonio Troyano Jiménez, Francisco Javier Ortega Rodríguez, Eugenio Martínez Cámara (eds.): Actas de las V Jornadas TIMM, Cazalla de la Sierra, España, 12-JUN-2014, publicadas en http://ceur-ws.org

- Phonetic sequence classification. Once the spoken utterance is phonetically transcribed, this sequence must be classified in order to determine the language of the utterance. A language model of sequences of phonetic units is learned for each language. The selection criterion if based on minimize the perplexity.

  Let $\mathcal{L}$ be the set of languages, $l_i \in \mathcal{L}$ one of this languages, and s the phonetic unit sequence to classify. The selected language $\hat{l}$ is the one that minimize the expression:

$$\hat{l} = \underset{l_i \in \mathcal{L}}{\operatorname{argmin}} \, 10^{-\frac{1}{|s|} \log p(s|l_i)} \tag{1}$$

  where, $p(s|l_i)$ is the probability of the sequence s assigned by the model representing language l_i.

3 Resources and Experimentation

This sections describes the resources used, how we learned the language models, and the preliminary experimentation carried out in this work.

3.1 Description of the used corpus

We have used a corpus of 3446 spoken sentences to learn the language models and evaluate our proposal. The sentences were uttered by several native English, French, and Spanish speakers. The distribution of the languages in the corpus was a little unbalanced (1338 in English, 708 in French, and 1400 for Spanish). The domain of the English and French sentences was queries to a information service about timetable and prices of long distance trains. The Spanish sentences were extracted from a unrestricted phonetically balanced corpus.

3.2 Learning the models

As phonetic unit, we have chosen context-dependent phonemes. Specifically, we have used triphones, ie, phonemes with information about the phonemes that appear to their left and right. We have learned the acoustic models for triphones and the models of sequences of triphones using an independent Spanish corpus. Only triphones for Spanish have been considered in this work. We have used the same set of Spanish triphones for all the experimentation.

We have phonetically transcribed all sentences in the corpus using two different Acoustic-Phonetic Decoding modules. In both modules the set of triphones and the acoustics models associated to them were the same; the difference was the model of sequences of triphones used as language model. The first APD module used a trigram model of sequences of triphones. To avoid the bias of using for all languages a trigram model of sequence of phonetic units (triphones) learned with Spanish corpus, a second module was learned using an equiprobable unigram model of triphones. This way, all sequences of phonetic units have the same a priori probability. As result, we got six phonetically transcribed utterances sets, two for each considered language using our two different APD modules.

3.3 Experimentation

In order to conduct the evaluation of our approach, we split the available corpus by language and use 80% for training the classification models, leaving the remaining 20% to evaluate the performance of the system. Since we have two possible different APD modules (trigrams and equiprobable unigrams), we were able to learn two set of language models. For each set, we learned an trigram language model for every language we are trying to discriminate.

We used SRILM Toolkit [Sto02] to estimated the phonetic language models of the classifiers and HTK Speech Recognition Toolkit [You06] to perform the phonetic transcriptions.

Two different experiments were conducted. The first experiment consisted of measuring the perplexity of the test sets. Table 1 shows the perplexity for all training and test combinations. Each column corresponds to the test set for a different language and using an specific APD module (*Trigrams APD* for the APD based on trigrams of phonetic units and *Equiprobable APD* for the APD based on equiprobable unigrams of phonetic units). In addition, each row corresponds to a classifier learned using the transcriptions of the training sentences of an specific language using an specific APD module.

As expected, Table 1 shows a lower perplexity for combinations where the language of the classifier and the language of test are the same. Regarding the APD module, lower perplexity occur when an APD based on

Table 1: Perplexity of the phonetic language models

		Test set					
		Trigrams APD			*Equiprobable APD*		
		French	English	Spanish	French	English	Spanish
Trigrams APD	French	8.24	11.62	12.16	27.94	33.07	19.86
	English	10.79	6.63	11.29	40.78	18.86	18.76
	Spanish	11.27	10.86	7.57	59.43	39.22	13.98
Equiprobable APD	French	12.06	14.89	17.07	15.64	19.17	19.05
	English	14.41	8.79	15.13	21.19	10.57	17.43
	Spanish	11.57	10.97	8.43	28.53	21.42	10.98

trigrams is used to transcribe the sentences, specially those in the test set. It seems that, the use of trigrams of phonetic units learned using a corpus only Spanish is not as critic as we a priori expected.

A second experimentation was conducted in order to evaluate the performance of the Language Identification system. The global accuracy of the system was 0.841 when *Trigram APD* module was used and 0.775 when *Equiprobable APD* module was used. As in the case of perplexity, the best accuracy result is obtained using the *Trigram APD* module. Table 2 shows the accuracy considering the different languages involved. The best results are obtained for Spanish, possibly because the triphones used were just those of Spanish. Although the phonetic similarity between Spanish and French seems bigger than the phonetic similarity between Spanish and English, results for English are better than those obtained for French. This may be due to the greater amount of English sentences available for the experimentation.

Table 2: Accuracy of the Language Identification system

	French	English	Spanish
Trigrams APD	0.793	0.850	0.960
Equiprobable APD	0.771	0.857	0.928

4 Conclusions and future work

In this paper we have presented a preliminary approach to the language identification problem. Our proposal is based on the classification of sequences of phonemes assuming that each language has its own phonotactic characteristics. The experimentation shows that our approach is able to predict reasonably well the language of the speaker, especially considering the limited resources used. We have many ideas on how to improve the performance of our system, including but not limited to using really language-independent phonetic units, using the recognizer lattices as input to the classification system.

Acknowledgements

This work is partially supported by the Spanish MICINN under contract TIN2011-28169-C05-01, Spain.

References

[Pal13] Palacios, C.S., D'Haro, L.F., de Córdoba, R., Caraballo, M.A.: Incorporación de n-gramas discriminativos para mejorar un reconocedor de idioma fonotáctico basado en i-vectores. Procesamiento del Lenguaje Natural **51** (2013) 145–152

[Rod13] Rodríguez-Fuentes, L.J., Brümmer, N., Peñagarikano, M., Varona, A., Bordel, G., Díez, M.: The albayzin 2012 language recognition evaluation. In Bimbot, F., Cerisara, C., Fougeron, C., Gravier, G., Lamel, L., Pellegrino, F., Perrier, P., eds.: Interspeech, ISCA (2013) 1497–1501

[Sto02] Stolcke, A.: Srilm - an extensible language modeling toolkit. In: Proc. of Intl. Conf. on Spoken Language. (2002) 901–904

[You06] Young, S.J., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.C.: The HTK Book, version 3.4. Cambridge University Engineering Department, Cambridge, UK (2006)