Gaussian Segmentation and Tokenization for Low Cost Language Identification

Ana Montalvo, José Ramón Calvo de Lara, and Gabriel Hernández-Sierra

Advanced Technologies Application Center (CENATAV),
7th Street #21416 between 214 and 216, Playa, Havana, Cuba
{amontalvo,jcalvo,gsierra}@cenatav.co.cu
<http://www.cenatav.co.cu/>

Abstract. Most common approaches to phonotactic language recognition deal with phone decoders as tokenizers. However, units that are not linked to phonetic definitions can be more universals, and therefore conceptually easier to adopt. It is assumed that the overall sound characteristics of all spoken languages can be covered by a broad collection of acoustic units, which can be characterized by acoustic segments. In this paper, such acoustic units, highly desirables for a more general language characterization, are delimited and clustered using Gaussian Mixture Model. A new segmentation method on acoustic units of the speech is proposed for later Gaussian modelling, looking for substitute the phonetic recognizer. This tokenizer is trained over untranscribed data, and it precedes the statistical language modeling phase.

Keywords: Spoken language recognition, Gaussian tokenization, acoustic segment modeling.

1 Introduction

Speech is the acoustic manifestation of the language, and probably the main communication way between humans. The development of telecommunications and information's digital processing has demanded efforts to understand the mechanisms of communication by speech. Among the numerous applications included in the field of speech signal analysis, is the process of detecting the presence of a given spoken language in a segment of speech by an unknown speaker, commonly referred to as spoken language recognition (LRE)[1].

It is known that humans recognize languages through a perceptual process that is inherent to the auditory system [2]. The perceptual cues that human listeners use, are an important source of inspiration for automatic spoken language recognition.

Many sources of information are imprinted on the speech signal, however those related with LRE task can be grouped in two broad classes, based on prelexical information and based on lexical semantic knowledge [3]. There are experiments which have proved that both, prelexical and lexical semantic knowledge contribute to the human perceptual process for spoken LRE.

Fig. 1. Levels of perceptual cues used for language recognition

Many representations of those desired cues, that distinctively characterize the different spoken languages, have been explored in the past few decades. Articulatory parameters [1], acoustic features [4], prosody [5] and phonotactic [6] are some of the features explored.

Although higher levels, like lexical and grammatical, contain much more language information than the other levels (see Fig.1) and tend to be more robust to the effects of channel distortion, they have the disadvantage of being more difficult to generalize (speech recognizers are required for each language to identify it, which requires transcribed audio so difficult to obtain) and can be computationally expensive. This makes the systems based on these standards rarely used in practice [2].

By contrast, the lower levels like the acoustic and prosodic ones, have the advantage of being easy to apply to any number of languages to a relatively low computational cost, but suffer from a lower precision and less robustness to channel variations and noise. On the other hand the intermediate levels like the phonotactic, maintains a balance between robustness, generalization and computational cost. While the acoustic information covers the physical sound patterns, the phonotactic refers to the constraints that determine possible syllables structures in a language.

Nowadays it is possible to distinguish two main groups of techniques for automatic LRE: acoustic level techniques and phonotactic level techniques. The phonotactic information combined with the acoustic information, are the most used features in LRE [7], being the ones based on phonotactics, those who provide the best compromise between the level of prior information needed for training the system and recognition accuracy. Nevertheless, currently, are merged both paradigms in most of commercial systems.

1.1 Problem Description

The phonotactic based systems use observed phono sequences to construct a statistical language model (LM) for each language of interest. This approach have two major drawbacks. Firstly, most of this systems use phonemes, a knowledge-based linguistic concept that is language dependent and in many cases difficult to deal with in speech processing. This systems require one or more phoneme recognizers, and each recognizer requires labeled data that are not always available (especially for rarely observed languages).

Secondly, the decoding phase of all these phonotactic approaches is computationally expensive, particularly when several recognizers in different languages run in parallel as in Parallel Phone recognition followed by LM (PPRLM) [6]. A model of acoustic units, language independent, can solve some of the just mentioned problems.

In this paper we propose to model certain acoustic segments of the speech signal, and to use them for later statistical modelling of each language. In other words, we will substitute the phoneme recognizer of a PRLM system, by a Gaussian tokenizer, trained with acoustic units, obtained through a data-driven acoustic unit recognizer. This way we eliminate the need of phonetically transcribed data for training, which allows that the tokenizer could be trained on the same acoustic data as that used for the LRE task, thus minimizing any session mismatch. With this method it is easier to increase the number of tokenizers since phonetically transcribed data are not required.

This investigation is focused on the main part of acoustic segmentation techniques, directed towards the LRE. In particular, this represents an incipient attempt to find an alternative representation to phonemes, which is a faintly studied topic today.

There are many other related issues, like channel normalization, channel adaptation, robustness under noisy conditions, but these issues are not investigated this time, they remain for future works.

This paper is organized as follows: Section 2 describes our main contribution which is the segmentation technique of the speech signal. In Section 3 and 4 is explained how the representation chosen was obtained. Then we described in Section 5 the results of the proposed approach, ending in Section 6 with conclusions and future works.

2 Speech Segmentation

As we said, many sources of information besides the linguistic message are carried on the speech signal. For the particular task of LRE a question that arises is whether phonemes, or other similar linguistically defined units are really needed. The underlying concept of words formed by phonemes may not be necessary. Over this ideas new approaches for LRE started to emerge.

In [9] Torres-Carrasquillo uses a sequence of Gaussian index to model the language information. Adami in [5] uses temporal trajectories of fundamental frequency and short-term energy to segment and label the speech signal into a small set of discrete units to characterize language. More recently Spada in [10] tried to approximate a phonetic segmentation using the variations in the spectrogram of the speech signal.

Our research contribute on this direction, trying to develop a technique for conditioning the Gaussian Mixture Model (GMM) tokenizer input. The idea was to eliminate the noisy token sequences and give prominence to longer duration events. We used information in the cepstral domain to segment the speech signal, supported by the GMM probability distribution of the feature vectors.

Fig. 2. Proposed Methodology: three first steps from training phase and the 4th from the test phase

2.1 Proposed Methodology

The features vectors can be seen as points in an N-dimensional feature space, where N is the dimension of feature vectors. Together with the influence of the variability session effects, these feature vectors represent also the state of our articulation organs. As the movements of our articulation organs are slow, it could be assumed that consecutive features in time domain, will be near in

cepstral domain too. In other words, acoustically stable sounds intervals, correspond to consecutive features in cepstral domain.

These ideas motivated us to think that a good way to define acoustic units could be by grouping together nearby features in the cepstral domain. Spada in [10] obtained acoustic segments using a spectral variation function based on the euclidean distance between the static MFCC to the left and the right of the current frame. Our proposal incorporates the dynamic of the features, adding the delta and double-delta information to the analysis. Not only looking for spectrally stable segments, but a more comprehensive pattern.

The first step is the creation of a Universal Background Model (UBM)[12] as is shown in Fig. 2. This model will be called the “slicer” UBM, because once obtained it will indicate the boundaries of the segments. We grant the quality of universal to this model, because it is independent of language and seeks to represent as much as possible acoustic phones, which is why you train it with all the languages that will be involved in the task.

One of the attractive attributes of the GMM is its ability to closely approximate any arbitrarily shaped data distributions, and its ability to model the underlying data classes by the individual Gaussian components. Here this GMM is used to approximate the overall acoustic-phonetic distributions of the languages over the model was trained on.

The trainset is clustered with this model, and the two most likely Gaussians for each frame are analyzed. Two frames will stay together if they share one of their two most probable Gaussians, otherwise they’ll be separated, and be part of different acoustic units. If there is a frame whose most probable Gaussians have no relationship with their neighbors, then that frame is removed. For the features corresponding to a same acoustic unit, we compute the average, in this way the whole segment is represented by an average single vector.

With this, we expect to reduce the cost of the systems in terms of resources and computational complexity, without compromising its discriminative power.

3 GMM for Voice Tokenization

After the segmentation is done, another Gaussian cluster is trained, but this time with much less classes. This model will be called the tokenizer UBM, and its function is to represent those acoustic units previously segmented and represented by an average vector, with the index of the most likely Gaussian of the tokenizer just like Torres-Carrasquillo did in [9]. With this approach, the vocabulary or number of tokens in the alphabet, is the same as the number of Gaussians in the UBM tokenizer. For our experiments we use 128 Gaussians clusters.

This aspect will be studied looking for an optimization in future works, because we are tempted to think that exist a relation between the number of clusters and the number of common phones of the languages that participate in the model creation.

4 Language Modeling

The purpose of the LM is to provide a syntax that defines possible tokens sequences and enables the computation of the probability ($P(W|L)$) of the token string $W = (w_1, w_2, \dots, w_Q)$ given the LM L .

The statistical LM is created using the training text set, which was generated by means of the previously trained cluster (UBM tokenizer). Making use of the Carnegie Mellon University Statistical Language Modeling toolkit [13] we obtained trigram models for each target language. At the testing phase, to estimate the probabilities of an index triplet (3-gram) we compute this quantity as:

$$P(w_i|w_{i-1}, w_{i-2}) = \frac{C(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})} \tag{1}$$

where $C(w_{i-2}, w_{i-1}, w_i)$ and $C(w_{i-2}, w_{i-1})$ are the frequency count of the triplet (w_{i-2}, w_{i-1}, w_i) and the index duplet (w_{i-2}, w_{i-1}) respectively, that occurred in the training set.

Thus for every string of Gaussian indexes, the probability of it is calculated as:

$$\log P(W|L) = \sum_{i=1}^Q \log P_L(w_i|w_{i-1}, w_{i-2}) \tag{2}$$

The language corresponding to the LM that maximizes $P(W|L)$ is selected as the language sought.

5 Experimental Results

The speech signal is divided into overlapping frames, 25 ms length with 10 ms frame shift, where it is supposed to be stationary. Mel Frequency Cepstral Coefficients (MFCC)[11], deltas (Δ) and double-deltas ($\Delta\Delta$) are used to describe each frame.

The experiments were conducted using a part of the Oregon Graduate Institute Multi-Language Telephone Speech (OGI) corpus [14].

A silences detector process took place before the Gaussian segmentation step. We use a trigram model to describe each language and the slicer UBM is estimated from the training selected part of the OGI corpus.

The selected target languages were english and spanish, and the training and test sets where defined as follows:

Table 1. Experiment Corpus

Training set			Test set (30s each signal)	
Language	Time	Number of signals	Language	Number of signals
English	1h 23min	102	English	109
Spanish	1h 23min	109	Spanish	45

It is worth noting the small volume of training samples (less than 3 hours) and the fact that we haven't made use of any phonetically transcribed data for the experiments.

The performance measure used for this task is the equal error rate (EER). It represents the system performance when the false acceptance probability (detecting the wrong language for a given test segment) is equal to the missed detection probability (rejecting the correct language).

Table 2. Performance comparison between our Gaussian segmentation based criterion and the Euclidean distance based criterion [10]

Language	EER Gaussian	EER Euclidean
English	22.01%	36.69%
Spanish	24.44%	36.66%

The results are compared using the 30-second test segments.

6 Conclusions

It's been presented a method for segmenting the speech signal into a number of acoustic units, seeking to find a more compact and language-independent representation. Of course the goal was to do this without a significant damage to the performance of the system, and the experimental results have shown promising values of EER for this approach.

For 30s test segments, results were around 23% of EER, which although still far from the results using phonetic recognition is a significant improvement to the proposed made in [10], much more appealing with the small amount of training data which is a limitation very commonly found in real world applications.

Futures studies will be conducted over the idea of increasing the number of tokenizers, looking for analogies with PPRLM, and the selection of more optimal features for the particular task. Also are objectives to incorporate more samples and more languages to the experiments. The aspect of the number of Gaussians clusters used in the GMM tokenizer is also pending of further insights, because as we said there is probably a relation between the number of clusters and the number of common phones of the languages that participate in the model creation.

References

1. Siniscalchi, S., Reed, J., Svendsen, T., Lee, C.: Universal attribute characterization of spoken languages for automatic spoken language recognition. *J. Computer Speech & Language* 27, 209–227 (2013)
2. Li, H., Ma, B., Lee, K.: Spoken Language Recognition: From Fundamentals to Practice. *J. Proceedings of the IEEE* 101, 1136–1159 (2013)

3. Zhao, J., Shu, H., Zhang, L., Wang, X., Gong, Q., Li, P.: Cortical competition during language discrimination. *J. NeuroImage* 43, 624–633 (2008)
4. Sugiyama, M.: Automatic language recognition using acoustic features. In: *Proc. ICASSP*, pp. 813–816 (1991)
5. Adami, A., Hermansky, H.: Segmentation of speech for speaker and language recognition. In: *INTERSPEECH 2003* (2003)
6. Zissman, M.: Comparison of Four Approaches to Automatic Language Identification of Telephone Speech. *IEEE Transactions on Speech and Audio Processing Journal* 4(1), 31–44 (1996)
7. Kempton, T.: *Machine-assisted Phonemic Analysis*. University of Sheffield (2012)
8. Muthusamy, Y., Jain, N., Cole, R.: Perceptual Benchmarks for Automatic Language Identification. In: *International Conference on Speech and Signal Processing*, pp. 333–336 (1994)
9. Torres-Carrasquillo, P., Reynolds, D., Deller, J.: Language identification using Gaussian mixture model tokenization. In: *ICASSP*, pp. 757–760 (2002)
10. Spada, D., López, I., Toledano, D., González, J.: Acoustic Event Recognition for Low Cost Language Identification. In: *V Jornadas en Tecnologías del Habla*, pp. 25–28. UAM (2007)
11. Davis, S., Mermelstein, P.: Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustic, Speech and Signal Processing* 28, 357–366 (1980)
12. Montalvo, A., Calvo, J.: *Métodos para reducir la variabilidad de sesión en el reconocimiento del locutor*. Technical report, CENATAV (2012)
13. Rosenfeld, R.: The CMU Statistical Language Modeling Toolkit and its use in the 1994 ARPA CSR Evaluation. *ARPA SLT 95* (1995)
14. Muthusamy, Y., Cole, R., Oshika, B.: The Ogi Multi-Language Telephone Speech Corpus. In: *ICSLP*, pp. 895–898 (1992)