Unsupervised identification of speech segments using kernel methods for clustering

José Anibal Arias

Equipe SAMOVA – IRIT/UPS
118, route de Narbonne, 31062 Toulouse cedex 04, FRANCE
arias@irit.fr

Abstract

Kernel k-means and spectral clustering have been used to separate input space clusters by means of non-linear mappings. In this paper we adapt and extend these methods to identify constitutive units of speech: consonants, vowels and silences. The discover of this structure is very useful for prosody-based systems of automatic language identification or language disorders detection. In order to find stable speech segments, infra-phonetic segmentation is performed using the divergence forward-backward algorithm. Our test corpus is a six-languages subset of OGI_MLTS corpus. We present better classification results than traditional approaches as well as faster processing times.

1. Introduction

Among the most important features to identify a language, we can find phonetic and phonotactic structure. Nevertheless, prosody (and particularly rhythm) can be also exploited to give clues about the "linguistic distance" between languages [1]. The advantage of rhythm information is that it can be extracted from a vocalic model of speech in a non-supervised way and without need of training corpus [2]. Our objective in this work is to provide a precise model of speech components that could be used for language characterization purposes.

Clustering is a fundamental problem of data mining. Clustering methods try to find data similarities in order to discover information. They can be used as a first step of more complex strategies of data analysis. In this paper we construct clusters from fundamental speech units. We use the divergence forward-backward (dfb) algorithm [3] to detect stable segments of speech that are clustered with kernels methods and identified according to its energy content. When we cluster speech segments in a two-class model, we get an excellent speech activity detector; when we define a three-class model, we have automatically a vowel-consonant-silence classifier.

The organization of the paper is the following: in section two we discuss the importance of vowel detection systems for language identification tasks and we explain the baseline system which serves as reference. In section three we explain our kernel k-means and spectral clustering approach based on segments issued from the dfb algorithm. In section four we describe the test corpus and present various results. Finally, in section five we present conclusions as well as future work challenges.

2. Baseline system

2.1. Vocalic systems

A typical language identification system consists of a phonotactic model of each language, trained with labeled corpus. Nevertheless, other research axes create unsupervised language prosodic models and achieve good degree of discrimination. The most important advantage of prosodic models is their simplicity and the fact that they do not need a training or adaptation phase for each language. A study [4] shows different types of prosodic models based on the vocalic structure of languages. Vowels have simpler articulatory representation than consonants, they are uniformly configured as front/back and open/closed. The number and duration of vowels in speech sequences is a potential source of information about the language that is spoken in the sentences (fig 1).

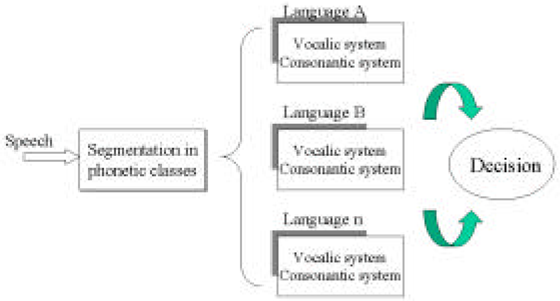

Figure 1: Language identification system based on vocalic system modeling.

2.2. Signal segmentation

One of the most difficult tasks in speech processing is to define limits of the phonetic units present in the signal. Phones are strongly co-articulated and there are no clear borders among them. Then, what is more facile is to detect stable zones in speech. The segmentation algorithm we used is the dfb: the speech signal is supposed to be a sequence of stationary and transitory zones; each one is characterized with a statistical auto-regressive Gaussian model. To define a border the dfb look at changes in vector coefficients of the model. In execution time, two models M_0 and M_1 are estimated from the acoustic signal. M_0 represents the segment from last break while M_1 is a short sliding window starting also after last break. When the Kullback divergence criteria change more than a certain limit, a new break is declared in the signal. The algorithm detects three sorts of segments: shorts or impulsive, transitory and quasi - stationary. In fig. 2 we show an example of dfb segmentation where infra-phonetic units are determined.

Segment's classification is performed in two steps [5]. First, a speech/silence decision is made based in the signal's standard deviation, and after, the energy is computed (weighted to give more importance to the spectral band 0.35-1kHz) to discriminate vowels from consonants.

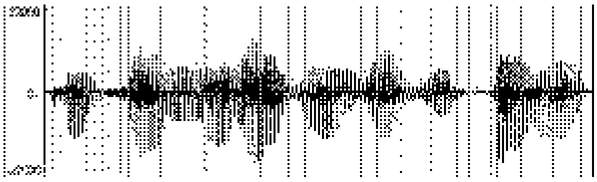

Figure 2: Automatic dfb segmentation of one-second speech (sentence in French).

3. Kernel clustering

Clustering creates subsets of data that are more uniform than the overall set. Ideally, all possible partitions of data should be tested to select the best one, but this is computationally infeasible [6]. With kernel methods we can efficiently construct clusters in kernel-defined space, taking advantage of the projection functions that allow non-linear separation of data in input space.

3.1. Kernel k-means

When the typical k-means clustering algorithm is enhanced by the use of kernel mappings, the objective function of clustering is defined as [7]:

$$D\left(\{p_j\}_{j=1}^{k}\right)=\sum_{j=1}^{k}\sum_{a\in p_j}\left\|f(a)-m_j\right\|^2 \qquad (1)$$

where p_j are the clusters and m_j is the "best" cluster representative:

$$m_j=\arg\min_z \sum_{a\in p_j}\left\|f(a)-z\right\|^2 \qquad (2)$$

The Euclidean distance from $f(a)$ to center m_j is:

$$\begin{aligned}\left\|f(a)-m_j\right\|^2 &= f(a)\cdot f(a)-\\ 2f(a)\cdot f(b)_{b\in p_j} &+ f(b)\cdot f(c)_{b,c\in p_j}\end{aligned} \qquad (3)$$

The dot products generated in (3) are computed using one kernel function. In the algorithm, we initialize the clusters randomly and we iterate until the objective function stops (monotonically) decreasing. Several pass are needed to find good cluster configurations.

This algorithm is prone to local minima since the optimisation is not convex (fig. 3), but can be improved adding prior knowledge at initialization.

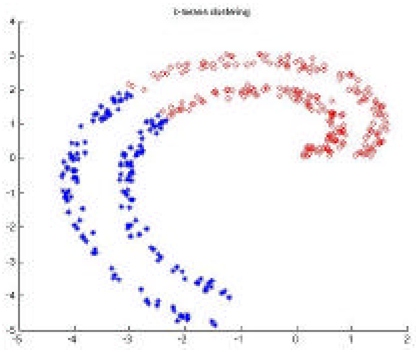

Figure 3: Example of two class kernel k-means clustering. This method do not work well with non-convex regions.

3.2. Spectral clustering

Spectral methods for clustering have emerged recently [8]. They use eigenvectors of a matrix derived from the distance between points to help determine the partitions. Several algorithms have been proposed using eigenvectors in different ways [9][10].

In [9] is proposed a spectral relaxation to solve the k-way normalized cut problem. An affinity matrix A (nonnegative and symmetric) is built to represent the distance between points. Let D be the diagonal matrix whose (*i*,*i*) entry is the sum of the entries of row *i* in matrix A. The top *k* eigenvectors of the matrix:

$$L = D^{-1/2} A D^{-1/2} \qquad (4)$$

are used to compute a discrete partitioning of the points.

The affinity matrix is defined by:

$$A_{ij} = \exp(-\|s_i - s_j\|^2 / 2s^2), \quad A_{ii} = 0 \qquad (5)$$

Spectral clustering give results very similar with what a human can choose (fig. 5), creating clusters that do not form convex regions or that overlap.

In order to compute kernel matrices, we decompose the speech signal in 5-seconds frames; otherwise memory requirements would be too important.

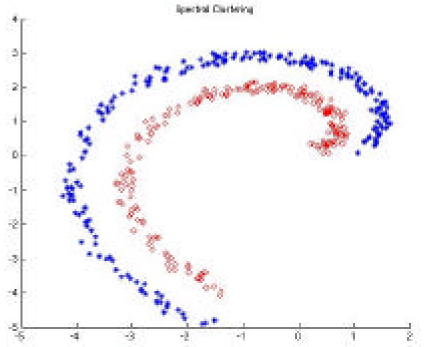

Figure 4: Example of two class spectral clustering. The topology of the clouds of points is well respected.

We use cluster segment's energy to label the resulting subsets: in a two class system, the cluster with lower mean energy is considered as the non-speech group. In a three class system, the previous rule is applied and we also consider that the cluster with higher segment's mean energy is the subset of vocalic segments.

4. Experiments

4.1. Test corpus

We use six-languages from the OGI_MLTS corpus to perform clustering tests: English, German, Hindi, Chinese, Japanese and Spanish. The corpus consist of telephonic spontaneous speech sampled to 8 kHz and presented in sentences of around 30 seconds length. The corpus has been phonetically labeled by experts following the CSLU rules [11]. We used 30 minutes of speech to perform tests.

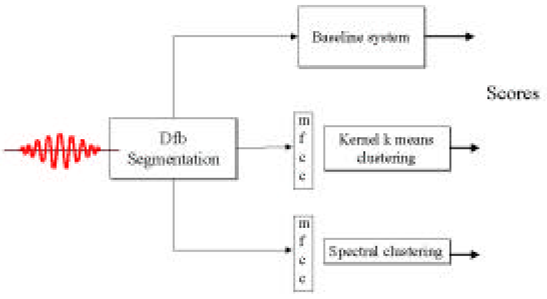

Figure 5: Experimental protocol to automatically detect speech components.

4.2. Pre-processing

We perform one feature extraction process to keep relevant characteristics from speech; we make cepstral analysis of data (fig. 5) [12]. The signal is decomposed in 16ms frames and for each frame 30 parameters are used: 14 MFCC plus energy and their associated derivatives. Cepstral features are normalized by cepstral subtractions. After automatic dfb segmentation, MFCC vector from the middle of the segment is taken to represent it entirely.

4.3. Clustering

The first test is a two-class clustering of speech components; as result we have a system that detects speech activity. The evaluation of results is made with one tool issued from NIST campaign's for audio indexing. This software calculates the accuracy of audio segmentation.

In two class clustering, the cluster with less energy correspond to the group of "silence", "pause" and "closure" etiquettes from manual labeling.

Table 1: Accuracy results of two-class automatic classification.

System	Baseline	Kk-means	Spectral
Speech detector accuracy	58.89%	86.54%	70.72%

The second test is a three-class clustering. We realize that this procedure identifies silence, vowels and consonants in speech with good accuracy and fast processing time (especially k-means). We process speech in 5-second frames, then kernel matrix compute the dot product of about 400-500 vectors per frame, which do not require high memory resources.

Table 2: Accuracy results of three-class automatic classification.

System	Baseline	Kk-means	Spectral
Vowel/consonant/silence detector accuracy	72.66%	72.89%	73.14%

In figure 6 we present an example of automatic identification of speech segments. The first row of etiquettes is the manual OGI labeling. The second row shows automatic dfb segmentation plus silence and vowel/consonant identification performed by the baseline system. The third row represents dfb segmentation plus three-class spectral clustering.

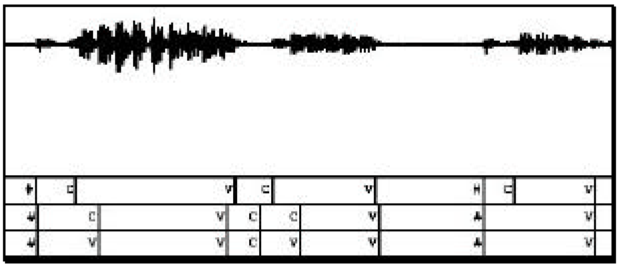

Figure 6: Example of automatic identification of speech segments.

5. Conclusions and future work

Clustering reveals as a promising alternative for speech processing. Our results are ready to be used by automatic language identification systems.

These are the first set of tests we perform to try to discover the complex speech structure in kernel feature space. It is obvious that the subsets of phones are strongly overlapped, but we will continue the research of more adapted kernel spaces to its representation.

Evolutions to clustering methods are waiting us: which prior information could be helpful for speech k-means clustering? How we can chose eigenvectors for better spectral clustering? Which kernel normalization will be more adapted to clustering? How many clusters we can create from speech signals?

We believe that increasing the number of clusters will allow us to identify more specific phones (fricatives, plosives, nasals) as well as different types of vowels. Other parameterization techniques should be explored together with clustering.

6. Acknowledgements

Thanks to Régine André-Obrecht and François Pellegrino for letting us use the algorithms of the baseline system. Thanks also to Jérome Farinas and Julien Pinquier for the opinions they give for conducting this research.
This work has received financial support from the Mexican National Council of Science and Technology.

7. References

[1] Farinas J., Rouas J. L., Pellegrino F. and André-Obrecht R. “Automatic extraction of prosodic features for automatic language identification”, To be published, 2005.

[2] Pellegrino, F., Farinas J. and Rouas J. L., “Automatic Estimation of Speaking Rate in Multilingual Spontaneous Speech”. *International Conference on Speech Prosody*, 2004.

[3] André-Obrecht, R., “A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals”, IEEE Trans. on ASSP, vol. 36 , n 1, 1988.

[4] Vallée N., *Systèmes vocaliques: de la typologie aux prédictions* Ph D Thesis, Stendhal University, Grenoble, France, 1994.

[5] Pellegrino, F., *Une approche phonétique en identification automatique des langues: la modélisation acoustique des systèmes vocaliques*, Ph D Thesis, Paul Sabatier University, Toulouse, France, 1998.

[6] Shawe-Taylor J. and Cristianini N., *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge, 2004.

[7] Dhillon I. S., Guan Y. and Kulis B., "Kernel k-means, Spectral Clustering and Normalized Cuts". *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2004.

[8] Weiss Y. "Segmentation using eigenvectors: A unifying view". *International Conference on Computer Vision*, 1999.

[9] Ng A. Y., Jordan M. I. and Weiss Y, "On Spectral Clustering: Analysis and an algorithm". *Neural Information Processing Systems (NIPS) 13*, 2001.

[10] Bach F. and Jordan M., "Learning Spectral Clustering". *Neural Information Processing Systems (NIPS) 16*, 2004.

[11] Lander T., *The CSLU Labeling Guide*, Center for Spoken Language Understanding, Oregon Graduate Institute, 1997.

[12]] Lu L., Zhang H. J. and Jiang H., “Content Analysis for Audio Classification and Segmentation “, IEEE Trans. on Speech and Audio Processing, vol. 10, n 7, 2002.