FROM VOCALIC DETECTION TO AUTOMATIC EMERGENCE OF VOWEL SYSTEMS

François PELLEGRINO and Régine ANDRE-OBRECHT

IRIT - 118, Route de Narbonne
F-31062 Toulouse Cedex - France
pellegri@irit.fr, obrecht@irit.fr

ABSTRACT

This paper presents our work on vowel system detection as part of a project of Automatic Language Identification using phonological typologies[1].

We have developed a vowel detection algorithm based on spectral analysis of the acoustic signal and requiring no learning stage. It has been tested with two telephone speech corpora:

- with a French corpus provided by the CNET, 7.4 % of detections are false while about 25 % of the vowels present in the signal are not found.
- experiments with 5 languages of the OGI_TS corpus [1] result in 88,1 % of correct detection and about 15 % of non-detection.

We also present in this paper the Vector Quantization (VQ) LBG-Rissanen Algorithm [2] that we use for vowel system modeling. Preliminary experiments are reported.

1. INTRODUCTION

At the approach of the XXIst century, world communication becomes an overwhelming reality and multilingual application developments are uprising. However, they require an automatic Language Identification (LId) system as front-end [3].

A wide range of distinctive features are available to characterize each language; they are present in several sources: acoustic, phonetics, phonology, morphology, prosody, syntax, etc. As in other speech processing applications, the challenge is to integrate the knowledge gathered by experts in an automatic system.

For the last decade, LId systems have been getting more accurate and the number of identified languages has increased [1,4].

Most of the systems are based on HMMs (Hidden Markov Models) and phonotactics (N-grams, etc.) [4, 5, 6]. The reverse of the medal is that extending an existing system to a new language needs a consistent amount of data and a long reestimation procedure.

Other systems take advantage of various features (Fo or Formants tracking...) through statistical modeling [7] or spectral distance calculation [8].

Recent phonological studies with the UPSID database [9] have resulted in a vowel system typology [10] and in improvements of vowel prediction models. According to this acoustic-based typology, a vowel system is characterized by the number of vowels, their position in an acoustico-articulatory space and the frequency of occurrence of the system in the UPSID database.

Exploiting this typology in an automatic LId system is an alternative and promising approach. We propose a validation strategy which consists in evaluating the opportunity to automatically get vowel system information from the speech signal through a vowel detection stage.

The 2nd Section presents the initial algorithm of vowel detection we implement and the results we get testing it with a French Telephone Speech corpus.

To verify that our method is independent of the tongue, we extend it to 5 languages from the OGI-TS corpus, and we propose some improvements. Section 3 describes this modified algorithm.

Section 4 deals with the problem of building a model of the vowel system out from the detected vowels. The VQ LBG-Rissanen is proposed and preliminary experiments are described.

2. FROM ACOUSTIC SIGNAL TO VOWEL SYSTEM

The strategy we propose to extract vowels from the raw speech waveforms is based on spectral analysis. It requires no learning and it is language independent.

2.1. Vowel Detection

Each signal frame is processed through a mel-scale filter bank resulting in a 24 energy coefficient vector.

A simple distance formula is applied to compute the *Sbec* criterion (Spectral Band Energy Cumulating):

$$Sbec = \sum_{i=1}^{24} \alpha_i \left| E_i(t) - \overline{E}(t) \right| \qquad (1)$$

where: - t is the number of the current frame

[1] This research is supported by the French "Ministère de la Défense" as part of an agreement with DRET.

- $E_i(t)$ is the energy in the ith Mel filter
- $\overline{E}(t)$ is the mean of filter energies
- α_i is the weight of the ith Mel filter

Generally speaking, a vowel is characterized by a high *Sbec* value, due to the presence of formants and gaps: maxima greater than an adaptive threshold are located and they correspond to potential vowels. In parallel, the "forward-backward divergence" algorithm [11] is performed to give a statistical segmentation of the signal, without *a priori* knowledge. The detected boundaries result in both short transient segments and long steady ones (as vocalic sections). An example of segmentation and *Sbec* calculation is given in Figure 1.

Each *Sbec* maximum is validated if the underlying segment duration is greater than 32 ms. This validation, based on both time and energy, enables to eliminate bursts and non significant segments.

2.2. Experiments

• ***The corpus***

To validate our approach we use a telephone speech corpus provided by the CNET[2]. It consists of twelve French words pronounced by 100 male and female speakers. 11 French vowels (including 2 nasal vowels) are present.

• ***The results***

The detection validation is based on an automatic segmental labeling developed at IRIT in a robust speech recognition task [12].

An example of detection is given in Figure 1, and Table 1a and Table 1b display the results on the CNET corpus.

More than 90 percent of the validated maxima are labeled as vowels; wrong detections are composed of bursts longer than 32 ms, fricatives ('s' for example) as well as vowels badly labeled because of a wrong alignment of the automatic labeling program.

About 25 % of the expected vowels are not detected. It mainly consists of 'i' and 'y' with low energy (maxima lower than the adaptative threshold).

3. TOWARDS MULTILINGUALITY

Since our research tends to identify languages, we test the *Sbec* criterion on a set of languages from the OGI_TS corpus. It appears that most errors consist of high energy unvoiced sounds (e.g. 'ʃ'). It leads us to develop a more accurate detection algorithm.

[2] The CNET is the French "Centre National d'Etudes en Télécommunications"

3.1. Improvement to Vowel Detection

The main flaw of the previous algorithm is that it is unable to eliminate maxima of *Sbec* that fit with unvoiced frames.

The new criterion, named Rec (Reduced Energies Cumulating) is :

$$Rec = \sum_{i=1}^{24} \alpha_i \left(E_i(t) - \overline{E}(t) \right)^+ \qquad (2)$$

where:
- t is the number of the current frame
- $E_i(t)$ is the energy in the ith Mel filter
- $\overline{E}(t)$ is the mean of filter energies
- α_i is the weight of the ith Mel filter

Unlike the first proposed algorithm, each *Rec* maximum is validated if two conditions are both verified :
- the underlying segment is longer than 15 ms,
- the distribution between low and high frequency energy must be balanced.

In fact, if we note :

Δt the duration of the underlying segment and

$$Rec = Rec_{LF} + Rec_{HF} \qquad (3)$$

where Rec_{LF} is the part of *Rec* corresponding to the Low Frequencies (300-1000 Hz),

and Rec_{HF} is the part of *Rec* corresponding to the High Frequencies (1000-3200 Hz),

maxima of *Rec* are validated if

$$\frac{Rec_{LF}}{Rec} \geq 0.5 \text{ and } \Delta t \geq 15 \text{ ms} \qquad (4)$$

Figure 2 gives an example of detection.

3.2. Experiments

• ***The corpus***

This new algorithm is tested with five languages (French, Japanese, Korean, Spanish and Vietnamese) from the OGI_TS corpus. Detections are checked using the broad phonetic labeling provided by OGI for about 25 speakers per language.

• ***The results***

Table 2a provides the number of correct detections, the number of wrong ones and the number of non-detected vowels, according to the hand-labeling.

Table 2b displays the percentage of vowels detected and the percentage of effective vowels in the set of detected ones according to the hand-labeling.

The results are homogenous among the 5 tongues and the *Rec*-based algorithm provides a better detection than the *Sbec* one: The number of detected vowels is higher (87 % instead of 75 % for French) with only a slight loss of quality (89.7 % instead of 92.6 % for French) although

the OGI_TS corpus is more difficult than the CNET one (spontaneous speech vs. isolated words).

4. FROM VOWELS TO VOCALIC SYSTEMS

4.1. Vowel system identification

To catch the vowel structure of the language, we propose to determinate how many patterns are necessary to correctly represent the structure of the vowels segments which have been gathered in the vowel detection stage.

Identifying a vowel system ignoring the number of vowels qualities is similar to build VQ codebook of unknown size. For that purpose, we propose a modified LBG algorithm based on both the classical LBG method coupled with a splitting algorithm [13], and on the Rissanen criterion [14]. The standard splitting LBG method is applied to the vowel segment set, and at each step, before splitting, we compute the following criterion:

$$I_n = -Ldg + 2n.p.\frac{\log N}{N} \qquad (5)$$

where : - Ldg is the log likelihood of the vowel set, when classing a codebook as a multigaussian distribution,
- p is the parameter space dimension,
- n is the number of codewords
- N is the cardinal of the vowel segment set.

Minimizing In results in the optimal number of codewords.

To implement this vector quantization, we compute for each vowel segment 8 MFCCs (Mel Frequencies Cepstral Coefficients) and we apply the LBG-Rissanen algorithm in the cepstral domain.

4.2. Preliminary Experiments

The data consist of the vowels detected in the CNET corpus and we study the quantization from two points of view:

1. Is the LBG-Rissanen VQ suitable for vowel quantization ?

2. What is its behavior if non vowel sounds are present among data ?

To answer the first question, we tested the VQ program with different sub-vowel systems derived from the detected vowels, i.e. we performed VQ with a different number of vowels in the data set: using the fundamental vowels 'i', 'a' and 'u' results really in a 3 words codebook; when extending the number of vowels qualities, the codebook size increases to a maximum value of 8 clusters. The Rissanen Criterion behaves adequately: an increase of data does not systematically result in an increment of the codebook size. However, the superposition of the mismatching vowel spaces of the 100 male and female speakers overcrowd the acoustic space: the increase of the codebook size would not be correlated with a significant gain of information.

To study the robustness of the LBG-Rissanen algorithm, we define two data sets derived from the whole data set: A first set consists of all the detections (correct detections and false alarms); it is named *Global Corpus*. The second set corresponds to the correct detections (only the segments labeled as vowels) and its name is *Clean Corpus*.

LBG-Rissanen VQ algorithm provides a 8 word codebook for both corpora. Figure 2 displays the resulting codebooks in the 2D principal space computed by Principal Component Analysis. The false samples do not result in important changes (given that the 2nd and 3rd clusters have permuted each other), and the VQ algorithm is quite robust to this noise.

5. CONCLUSION

This work proves that it is possible to extract vowel system information from the acoustic signal. Our present purpose is to improve the LBG-Rissanen modeling with a statistical normalization. Introducing phonological knowledge in a multilingual context (OGI_TS corpus) is the next stage towards Language Identification.

Correct Detections	Wrong Detections	Non detected vowels
2507	199	803

Table 1a: Results of the vowel detection with the CNET data.

Number of Detections	% of effective vowels	% of detected vowels
2706	92.6	75.7

Table 1b: Results of the vowel detection with the CNET data - Accuracy Rates

Language	Correct Detections	Wrong Detections	Non detected vowels
French	930	107	137
Japanese	674	56	151
Korean	813	146	129
Spanish	873	83	168
Vietnamese	520	120	120
Whole Data	3810	512	714

Table 2a: Results of the vowel detection with the OGI data.

Language	Number of Detections	% of effective vowels	% of detected vowels
French	1037	89.7	87.2
Japanese	730	92.3	81.7
Korean	959	84.8	86.3
Spanish	956	91.3	93.5
Vietnamese	640	81.2	81.2
Whole Data	4322	88.1	84.2

Table 2b: Results of the vowel detection with the OGI data - Accuracy rates

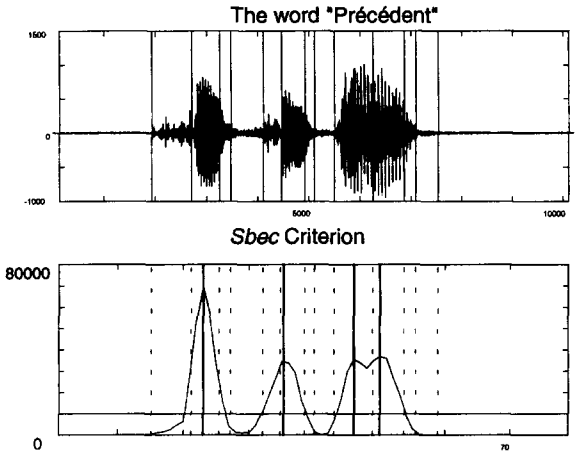

Figure 1: Example of Vowel Detection
a) Speech signal and statistical segmentation
b) *Sbec* and detected vowels (vertical solid lines)

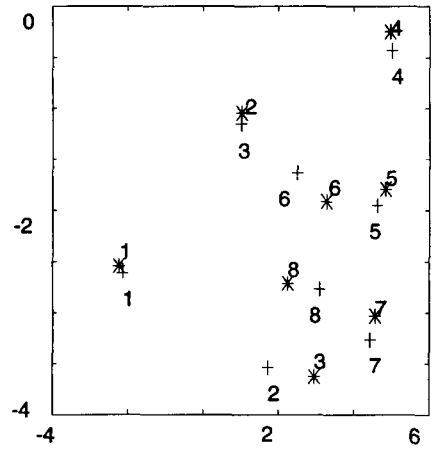

Figure 3: Result of PCA with 2 codebooks
'*' → *Clean corpus* VQ codebook
'+' → *Global corpus* VQ codebook

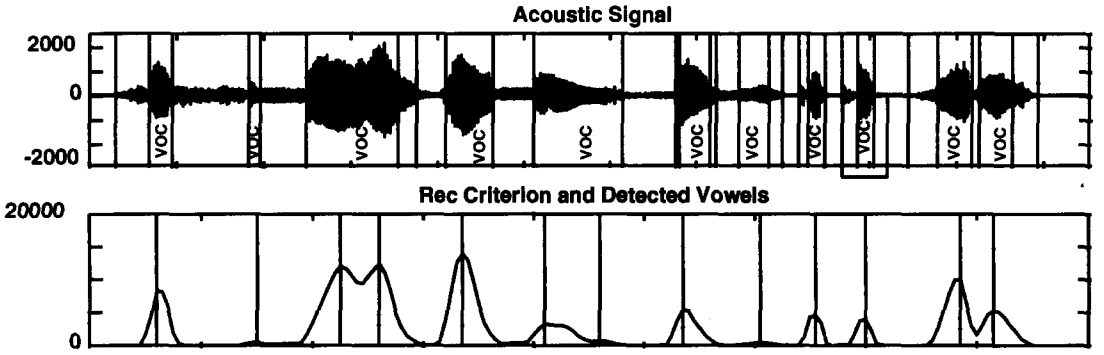

Figure 2: Example of Vowel Detection
a) Speech signal and hand vowel labeling
b) *Rec* and detected vowels (vertical solid lines)

"Je suis né à Guernon dans une petite ville"
"ʒ ə s ɥ i n e a g ɛ r n ɔ̃ d ã z y n p ə t i t v i l ə"

6. REFERENCES

[1] T. L. Lander, R. A. Cole, B. Oshika, M. Noel, "*The OGI 22 Language Telephone Speech Corpus*", Eurospeech 95, Madrid, pp. 817-820

[2] R. André-Obrecht, "*Segmentation et Parole ?*", Habilitation à diriger des recherches, Université de Rennes, IRISA, June 1993

[3] Y. K. Muthusamy, E. Barnard, R. A. Cole "*Reviewing Automatic Language Identification*" IEEE Signal Processing Magazine 10/94, pp. 33-41

[4] M.A. Zissman, "*Comparison of Four Approaches to Automatic Language Identification of Telephone Speech*", IEEE Trans. on SAP, Jan. 1996, Vol. 4, No 1, pp. 31-44

[5] Y. Yan, E. Barnard, "*An Approach to Language Identification with Enhanced Language Model*" Eurospeech '95, Madrid, pp. 1351-1354

[6] T. J. Hazen, V. W. Zue, "*Recent Improvements in an Approach to Segment-Based Automatic Language Identification*", ICSLP 94, Yokohama, pp. 1883-1886

[7] S. Itahashi, L. Du, "*Language Identification Based on Speech Fundamental Frequency*", Eurospeech '95 Madrid, pp. 1359-1362

[8] K. P. Li, "*Automatic Language Identification using Syllabic Spectral Features*", ICASSP 94 Adelaide, pp. I.297-I.300

[9] I. Maddieson, "*Patterns of Sounds*", Cambridge University Press, 1984

[10] N. Vallée, "*Systèmes vocaliques : de la typologie aux prédictions*", Thèse de Doctorat es Sciences du Langage, Université Stendhal, Grenoble, October 94

[11] R. André-Obrecht, "*A New Statistical Approach for Automatic Speech Segmentation*", IEEE Trans. on ASSP, Jan. 1988, vol. 36 no 1 pp. 29-40

[12] J.B. Puel, R. André-Obrecht, "*Robust Signal Preprocessing for HMM Speech Recognition in Adverse Condition*", ICSLP 94, Yokohama, pp. 259-262

[13] Y. Linde, A. Buzo, R. M. Gray, "*An Algorithm for Vector Quantizer Design*", IEEE Trans. on COM. Jan. 1980, vol. 28 pp. 84-95

[14] J. Rissanen, "*A Universal Prior for Integers and Estimation by Minimum Description Length*", The Annals of Statistics, 1983, Vol. 11, No 2, pp. 416-431