FROM LARYNGOGRAPHIC AND ACOUSTIC SIGNALS TO VOICING GESTURES

Nathalie Parlangeau, Régine André-Obrecht

Institut de Recherche en Informatique de Toulouse
118, Route de Narbonne
31062 Toulouse Cédex - France
parlange@irit.fr, obrecht@irit.fr

ABSTRACT

Many researchers have seen in the articulation an intermediate level of representation. In the gestural phonetic theory, units are articulatory gestures. In order to assess this theory with observed parameters, we have defined a robust labelling system (AMULET) of the multi sensor ACCOR speech database. Main articulatory gestures searched are Voice Onset and Voice Termination on both acoustic and laryngographic signals. We present here two efficient Voiced/Unvoiced/Silence detectors for the acoustic signals and a third one for the laryngographic signal.

1. INTRODUCTION

Speech production is a complex process relying on coordinated gestures, but the acoustic signal does not exactly depict its underlaying organisation. The question that leads up is : what is the right level of representation ?

Traditionally, the linguistic description of speech is strictly linear. Various approaches have been proposed attempted to formalise more enriched conceptions of phonological structure : incorporation of the syllable structure, explicit incorporation of consonant-vowel skeleton,... Even if these approaches have increased the range of facts that can be formalised in phonological theory, they do not explicit the relation between phonological and physical structure of speech. Many researchers have seen in the articulation an intermediate level of representation. Some perception studies have shown that the perception of linguistic structure is based on the articulatory structure of speech. What is perceived are articulatory gestures.

Based on these perception studies, the gestural phonetic theory is an alternative to previous ones like the Motor Theory, disproved as too simple. Browman and Goldstein [1] have abandoned the traditional vision of linguistic units as mental and abstract processes. They postulate that the linguistic organisation can be described with observable parameters. Lexical units are described in terms of articulatory gestures. A gesture is a basic action of the vocal tract (constriction or release) through space and over time. These spatio-temporal gestures are defined by specifying second-order dynamic equations.

In order to assess the gestural phonetic theory with observed articulatory parameters, we have defined the AMULET (Automatic MUltisensor speech Labelling and Event Tracking) system. It provides a robust annotation of the ACCOR multisensor speech database [2], in terms of articulatory events [3].

One of the main articulatory gestures is the vibration of the vocal cords, which is associated to the articulatory events : VO (Voicing Onset) and VT (Voicing Termination). These events may be detected on both laryngographic and acoustic wave. Nevertheless, on the acoustic signal, this information may be overlapped by a noise when a vocal tract constriction happens.

To explore this voicing gesture and the correlation between acoustic and articulatory events, we have developped and validated three Voiced/Unvoiced/Silence (VUS) detection systems. Two detectors are proposed for the acoustic signal and a third one is proposed for the laryngographic wave. Each detector is assessed by comparing results to a hand made labelling.

2. THE ACCOR DATABASE

The multisensor speech database was developped in the ESPRIT II Basic Research Action « ACCOR » (Articulatory acoustic Correlation of Coarticulation patterns). It includes articulatory and aerodynamic as well as acoustic data. Five signals were recorded simultaneously for each sentence : the acoustic signal, the vibrations of the vocal cords obtained by laryngography, the nasal and oral signals, binary images of 64 points representing the tongue contact with the palate : E.P.G.

3. AUTOMATIC ANNOTATION OF VOICING GESTURES

During the two last decades, many Voiced/Unvoiced/Silence detectors have been studied to be used in speech analysis/synthesis, coding or recognition systems. The most currently used parameters are energy, zero-crossing rate, autocorrelation, reflexion coefficients,... and the classification methods are based on weighted distance measurements [4], change

detections [5], neural networks [6] or Hidden Markov Models.

3.1. The acoustic Voiced/Unvoiced/Silence detection

Our two acoustic detectors are based on an *a priori* segmentation of the signal, the Forward-Backward divergence algorithm [7] : the signal is assumed to be a sequence of stationary units, each one is characterised by an autoregressive model Θ (LPC). The divergence test is based on the monitoring of a suitable statistic distance between two models Θ1 and Θ2. The procedure is performed in parallel on the signal as well as on the high pass filtered signal. To avoid omissions, if the delay between two boundaries is too long, the signal is performed in backward direction. The parameters (AR order and threshold) are speaker independent, and trained on another database.

3.1.1. A rule based voicing test

The four classical parameters are extracted from each segment on a centered window :

- the energy (E),
- the autocorrelation (R1),
- the first reflection coefficient (RC1),
- and the zero-crossing rate (Z) .

A set of rules monitores E and R1 in order to give a first decision. This decision is corrected by RC1 and Z.
The set of rules was trained on a little subset of data, two repetitions of one sentence for two speakers.
This test is used in order to detect the VO and VT events. Each segment is labelled as voiced, unvoiced or silence. Consecutive voiced (resp. unvoiced or silence) segments are merged, and global boundaries give VO and VT events.

3.1.2. The Vector Quantization method

◊ *The LBG-Rissanen algorithm*

The baseline system we implement is a standard LBG splitting method [8]. The algorithm provides for each class (voiced V, unvoiced U and silence S) a specific codebook. We assume that each class (V,U and S) can be modeled by a multi gaussian distribution, where each elementary gaussian density function corresponds to a codeword.
It appears that optimizing the codebook size for each set of data may result in a more adequate. So, we use an information criterion calculated from the data set : the Rissanen criterion *I(n)*.
In the LBG-Rissanen algorithm [9], we compute the criterion before splitting. Minimizing *I(n)* results in the optimal number of references.

$$I(n) = -Ldg + 2\, n \,.\, p \,.\, \log N$$

where : - *Ldg* is the log likelihood of the training set, when classifying a codebook as a multi gaussian distribution
- *p* is the parameter space dimension,
- *n* is the number of codewords,
- *N* is the cardinal of the training set.

◊ *Decision rule*

We first gather the three codebooks. The multi dimensional reference map may be considered as a simple codebook, and the decision rule is the classical K-mean classification based on the Euclidean distance in the cepstral space.

◊ *Implementation*

All sentences and nonsense words of the training data set are hand labelled in terms of voiced/unvoiced/silence segments, and this labelling is projected on automatic segments.
On these segments we extract a eight MFCC vector, as well as log(E), RC, RC1 and Z.
Multiple codebooks are constructed in order to determine the best set of parameters. Experiences are described in the fourth section.

3.2. The laryngographic voiced/unvoiced detection

The VO and VT events detection is based on a simplified version of the Forward-Backward Divergence algorithm.
We interpret each segment as voiced (resp. unvoiced) using a voicing test based on an adaptative level crossing ratio.

3.2.1. The adaptation of the Forward-Backward algorithm

The Forward-Backward method is performed on null mean signals. As the laryngographic signal has a variable mean, we calculate sequentially the signal mean and we process the corrected mean. In this simplified version, detection on the high pass filtered signal is not included.

3.2.2. The adaptative voicing detection

Once changes are detected, we interpret each segment as voiced/unvoiced using a voicing test based on an adaptative level crossing ratio which is applied for each segment on a centered window.
We define two levels on both sides of the signal mean. We calculate the two level crossing rate. We observe an important rate for the voiced segments and a very low one for the unvoiced segments. So, this test is very robust.
The frontiers of the segments are interpreted as VO and VT events.

4. EVALUATION AND RESULTS

4.1. The rule based voicing test

To evaluate our test, two experiments are done. In the first we evaluate labelling of each segment in terms of V, U and S, and in the second one we evaluate the VO and VT articulatory events.

4.1.1. First experiment

The training corpus is composed of two repetitions of one sentence of two different speakers (male and female).
The evaluation corpus is composed of five repetitions of three sentences for two different female speakers, one speaker is part of the training set.
The evaluation is made comparing the projection of the hand labelling voiced (resp. unvoiced, silence) segments on the automatic segments to the automatic voicing detection. Table 1 is the confusion matrix between automatic and labelling detection. We perform both the recognition rate and the reliability rate.

		Hand labelling			
		S	U	V	**Reliability**
	S	2499	268	36	**89.15%**
Aut	U	235	530	236	**52.94%**
	V	76	84	1693	**91.34%**
Recognition		**88.9%**	**60%**	**86.15%**	

Table 1 : rule based detector. V/U/S recognition and reliability rates.

4.1.2. Second experiment

In this evaluation, we have measured the delay between the automatic and the hand labelling under 10 ms, between 10 ms and 20 ms as well as over 20 ms. We have also taken omissions (O) and insertions (I) into account.
Delays greater than 20 ms are often due to a persistent sinusoidal wave (table 2). Insertions of VO and VT events will be interpreted in the future, with supplementary treatments as consonantic areas detections.

	<10	10<20	>20	O	I
VO	69/73	1/73	3/73		2
VT	61/73	4/73	8/73		2

Table 2 :. rule based detector. VO and VT evaluation.

4.2. The Vector Quantization method

4.2.1. Corpus

A first experiment is speaker independent. The training data set is composed of five repetitions of two sentences for one speaker, and some nonsense words VCV (V : /a,i,u/ and C :/p,t,k,tS,st,kl/). The recognition data set is composed of five repetitions of a sentence (not included in the training set) for a different speaker (corpus1).
The second experiment is multi-speaker. The training data set is composed of five repetitions of two sentences for two speakers, and some nonsense words VCV (V : /a,i,u/ and C :/p,t,k,tS,st,kl/). The recognition data set is composed of five repetitions of five repetitions of one sentence for the two same speakers (corpus2).
To assess the LBG-Rissanen algorithm, we have performed five experiments with different parameters. We retain two families in this presentation :

- Set4 : 4 MFCC, log(E), RC1, Z,
- Set5 : log(E), RC1, Z.

4.2.2. Experiments

For corpus1 as well as for corpus2, we perform the LBG algorithm and the LBG-Rissanen algorithm for each set of parameters.

◊ ***First experiment*** : corpus1, set 4 & set 5.

To present our results, we give the confusion matrix and we indicate the number of codewords *n* for each elementary codebook V,U and S.

		Hand labelling			
		S	U	V	**Reliability**
	S *n=10*	168	19	3	**88%**
Aut	U *n=10*	14	17	22	**32%**
	V *n=8*	17	11	193	**87%**
Recognition		**84%**	**40%**	**89%**	

Table 3 : recognition and reliability rates.Set4, Corpus1.

		Hand labelling			
		S	U	V	**Reliability**
	S *n=17*	184	32	19	**78%**
Aut	U *n=4*	6	6	5	**35%**
	V *n=4*	9	9	194	**92%**
Recognition		**92%**	**14%**	**89%**	

Table 4 : recognition and reliability rates. Set4, Corpus1.

		Hand labelling			
		S	U	V	**Reliability**
	S *n=4*	74	1	0	**99%**
Aut	U *n=40*	119	43	58	**20%**
	V *n=4*	6	3	160	**95%**
Recognition		**37%**	**91%**	**73%**	

Table 5 : recognition and reliability rates. Set4, Corpus1.

We can observe that raising *n* does not increase globally neither the recognition rate nor the reliability rate.
But an imbalance between the cardinal n of the different classes has to be controlled. For example, if the cardinal is increased for the class U (table 5), S attracts U, the reliability rate decreases for U and the recognition rate increases ; the contrary is observed when raising the cardinal of the class S.
The introduction of the four MFCC globally increases the recognition rates, and the different reliability rates are balanced (table 6).

		Hand labelling			
		S	U	V	**Reliability**
	S *n=10*	167	22	8	**85%**
Aut	U *n=10*	15	19	5	**49%**
	V *n=10*	17	6	205	**90%**
Recognition		**84%**	**40%**	**94%**	

Table 6 : recognition and reliability rates. Set5, Corpus1.

◊ ***Second experiment*** : corpus 1, set 4.

		Hand labelling			
		S	U	V	**Reliability**
	S *n=10*	275	36	9	**85,9%**
Aut	U *n=10*	26	49	8	**59%**
	V *n=10*	28	16	379	**89.6%**
Rec. rate		**83.5%**	**48.5%**	**95.7%**	

Table 7 : recognition and reliability rates.Set4,Corpus2.

The best result is obtained for the multi speaker experiment, with Set4 (table 7).

As a conclusion, the less reliable class is the unvoiced class, we observe a confusion between the U and S classes.
In our study of VO and VT detection, the two V/U/S detectors can be considered as V/UV detectors. In this case, the recognition rate of the unvoiced class exceeds 93% and the reliability rate 95%.

4.3. The laryngographic detector

The evaluation procedure is the same as in 4.1.2. .
As shown in table 3, we observe good results. Delays greater than 10ms are due to a persistent sinusoidal wave.

	<10	10<20	>20	O	I
VO	59/61			2/61	1
VT	56/61	3/61		2/61	1

Table 9 : laryngeal VO and VT evaluation .

5. CORRELATION BETWEEN ACOUSTIC AND LARYNGOGRAPHIC VOICING GESTURES

In order to study the correlations between laryngographic and acoustic voicing gestures, we have projected the laryngographic voicing gestures on the hand labelling acoustic signal (table 9).

	Silence	Unvoiced	Voiced
Unvoiced		93.3%	6.6%
Voiced	1.6%	0.4%	98%

Table 9 : Confusion matrix between laryngographic and acoustic voicing activities.

The laryngographic voicing gestures are good indicators of the voicing activity on the acoustic signal. Most of the confusions reveals a persistent voicing wave. For instance, it concerns the production of voiced plosives sounds, here /g/ and /d/ (figure 1) : voice bar is the result of both a persistent voicing wave and an overlapping noise due to the vocal tract constriction. The voice bar is not the result of a laryngeal voicing activity. Statistical results on nonsense words (Vowel- Voiced Plosives-Vowel) are in process.
These observations confirm that the end of the vibrations on the acoustic signal does not always correspond to the end of a voicing gesture. Any automatic labelling procedure will give the true end of voicing gestures without a delay.

6. CONCLUSION

We have proposed three efficient, speaker-independent and robust methods of voiced/unvoiced/silence detection. These three methods are used to label the ACCOR multisensor speech database and to extract indicators of the voiced gesture. The experiments prove clearly their consistency.
Then we propose a preliminary study of the spatial temporal correlation between the voicing events of the laryngographic and these of the acoustic signal. The observed delays shows interesting features. Our automatic procedures will enable to label efficiently a large amount of data to collect more statistical significant results on the study of various correlations.

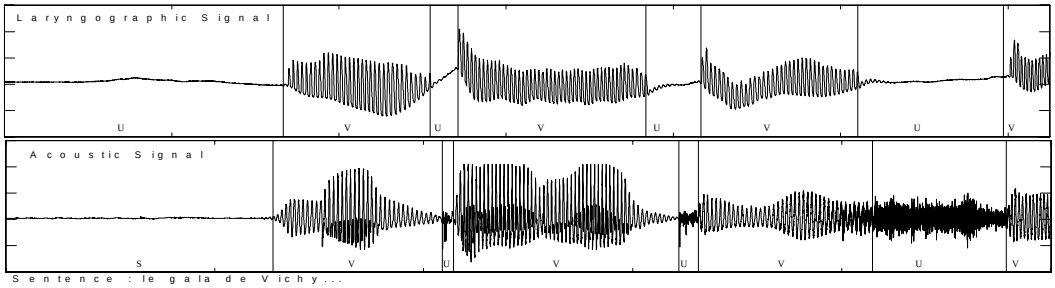

Figure 1 : correlation between laryngeal and acoustic voicing activity.

7. REFERENCES

[1] C.P. Browman, L.M. Goldstein, (1983) *«Towards an articulatory phonology »,* Phonology yearbook,, 1983, pp 219-252.
[2] A. Marchal, W.J. Hardcastle, (1993) *« ACCOR: Instrumentation and database for cross-language study of coarticulation »*, Language and Speech, 2-3, pp 137-153.
[3] N . Parlangeau, R. André-Obrecht, A. Marchal, (1996) *« Automatic articulatory annotation of multi-sensor speech database »*, ICASSP'96, Atlanta.
[4] B.S. Atal, L.R. Rabiner, (1976) *« A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition »*, IEEE Trans. On ASSP, vol. ASSP-24, No. 3, June 1976.
[5] C. Chan, (1986) *«Voiced-Unvoiced segmentation »,* ICASSP'86, vol. 1, pp 2271-2274, Tokyo.
[6] R.P. Cohn, (1991) *«Robust voiced-unvoiced classification using neural nets »,* ICASSP'91, vol. 1, pp 437-441, Toronto.
[7] R. André-Obrecht, (1988) *« A new approach for the automatic segmentation of continuous speech signals »,* IEEE. Trans on ASSP, 1988.
[8] Y. Linde, A. Buzo, , R.M. Gray, (1980) *« An algorithm for Vector Quantizer Design »,* IEEtrans. On COM. Jan. 1980, vol 28, pp 84-95.
[9] J. Rissanen, (1983) *« An universal prior for Integers and Estimation by Minimum Description Length »* , The Annals of Statistics, 1983 , Vol. 11 , No 2, pp 416-431.