AMULET : Automatic MUltisensor speech Labelling and Event Tracking : study of the spatio-temporal correlations in voiceless plosive production

Nathalie Parlangeau [*] *- Alain Marchal* [**]

[*] Université Paul Sabatier Institut de Recherche en Informatique de Toulouse
118, Route de Narbonne 31062 Toulouse Cedex

[**] Université d'Aix en Provence I URA 261 Parole et Langage
29, Ave R. Schuman 13621 Aix en Provence

ABSTRACT

Speech production is a complex process relying on coordinated gestures, but the acoustic signal does not depict its underlaying organization. Accepting that articulatory gestures are directly recognized through the coarticulation process, our proposal is to investigate the correlations between acoustic and articulatory informations and to assess gestural phonetic theory. We present here the framework for this investigation, the automatic articulatory labelling of the multi-sensor speech database ACCOR, and the study of the spatio-temporal correlations in the voiceless plosive production.

1. INTRODUCTION

To design Automatic Speech Recognition Systems, the main difficulty lies with the extremely large variability of the speech signal. This problem has been known and studied for a long time. One aspect is due to the assimilation and coarticulation phenomena : the assimilation is due to the phonological process whereas transitions between sounds are smoothed and phonetic features are spread over contiguous sounds. The coarticulation is inherent to the way speech is produced by the continuous motion of articulators [1]. Speech production is a complex process relying on coordinated gestures, but the acoustic signal does not immediately reflect the underlaying organization. The question that leads up is : what is the right level of representation ?

An hypothesis postulates that the articulatory gestures are directly recognized through the coarticulation process. From a theorical point of view, many researchers have seen in the articulation an intermediate level of representation which could link perception and production. The gestural phonetic theory is an alternative to previous theories like the motor theory which has been disproved as too simple [2]. Our proposal is to investigate the correlations between acoustic and articulatory informations in order to precise this intermediate level of representation and to assess the gestural theory. We first propose to study these correlations in the voiceless production process. This study will permit to define a robust identification system for voiceless plosives.

This work is performed on the multi-sensor speech database developed in the ESPRIT II Basic Research Action ACCOR (Articulatory Acoustic Correlations of Coarticulatory patterns) [3]. This database includes articulatory and aerodynamic as well as acoustic data. We dispose of five signals : the acoustic signal, the laryngograph trace, the nasal and oral airflow and the ElectroPalatoGraphic patterns (E.P.G.).

As far as voiceless plosive detection is concerned, we will only take into account the signals involved in the voiceless plosive production process. These signals are the oral airflow and the EletroPalatoGraphic patterns, as well as the acoustic signal. A complete presentation of AMULET can be found in [4].

In a first partof this manuscript, we present the theorical frame of this work. According to the gestural phonetic theory, we precise the nature of the different gestural units used. These gestures are based on articulatory events which are automatically detected by AMULET.

AMULET is a tool to detect automatically these articulatory events and we use it to label the ACCOR database. For each signal, we have developed automatic labelling procedures borrowed from Signal Processing in a precise methodological frame. These methods have been evaluated on a dataset of French sentences.

To conclude, we present the first study of the spatio-temporal correlations between acoustic and articulatory informations to precise the voiceless plosive production.

2. THE GESTURAL UNITS

2.1. The gestural phonetic theory

In the gestural phonetic theory, Browman and Goldstein [5] have abandonned the traditionnal vision of linguistic units as mental and abstract processes. They postulate that the linguistic organization can be described with observable parameters. So, they search speech invariance in the articulatory structure of speech. This is based on the perception study of linguistic structure by C. Fowler [6].

This theory describes the lexical units as articulatory gestures. A gesture is a basic action of the vocal tract during speech production. It consists in constriction and release of the vocal tract. Gestures are specified in terms of tract variables responsible of the constrictions. Each tract variable is associated to a set of articulators which movements determine the tract variable value. They form a coordinative structure. For example, a lip gesture is composed of two tract variables, based on the same coordinative structure :

Lip gesture = Lip Protrusion = (upper, lower lips, jaws)

Lip Aperture = (upper, lower lips, jaws)

2.2. Our gestural units

We dispose of four articulatory signals as well as the acoustic data. These five signals were recorded simultaneously for each sentence :

- the acoustic signal sampled at 20 kHz,
- the vibrations of the vocal cords obtained by laryngography (10 kHz),
- the nasal and the oral volume velocity trace (500 Hz),
- binary images of 8*8 points (1 image/5ms) representing the tongue contacts with the palate : E.P.G..

The gestural units are constrictions and releases of the vocal tract, observable on the four articulatory signals. A coordinative structure is associated with each gesture. A coordinative gesture is a set of articulatory events (table 2).

We present in the following table (table 1) the different gestural units and the articulatory events associated. We only describe the gestures for the two signals we are interested in for the voiceless plosive production process.

	Gestural units	**Articulatory events**
Oral	Weack Constriction	(BFO,MFO,mfO,DFO)
Airflow	Complete Constriction	(CCO,CRO)
EPG	Complete Constriction	(ACE,CCE,MCE,CRE,ind)

Table 1 : Gestural units. « ind » is not an articulatory event, but an index on the place of constriction, palatal or velar.

	Significance	Acoustic	Oral	E.P.G.
VO	Voice Onset	W		
VT	Voice Termination	W		
CC	Complete Constriction		O	E
CR	Constriction Release	W	O	E
BF	Build-up		O	
DF	Decline		O	
MF	Maximum		O	
mF	Minimum		O	
AC	Approach Closure			E
MC	Maximum Closure			E

Table 2 : Articulatory events

As far as acoustic data is the output of the articulation process, we do not want to detect articulatory gestures as such, but acoustic events revealing articulatory gestures (table 2).

3. AUTOMATIC LABELLING METHODS

3.1. Methodological frame

The annotation of the database is based on the following two principles :

- **non-linearity**,
- **channel-independency** of the informations.

The first principle is adopted to lead to proper annotation and not to preclude any a priori theorical assumptions about coarticulation. The methodological principle of channel-independency of the annotation is important to allow for the systematic investigation of the correlations between different channels of information. We have added a third one which is the robustness, in the sense that each labelling method has to be speaker-independent and that the detections must be consistent.

All labelling methods are built on the same schema : we first detect the discontinuities on the signal, and we interpret them as articulatory events. They are marked in the temporal domain according to precise criteria.

3.2. The Acoustic Signal

Automatic Segmentation Method

We first detect the acoustic discontinuities using a robust automatic segmentation method, the Forward-Backward divergence method [7] : the signal is assumed to be a sequence of stationary units, each one is characterized by an autoregressive model Θ (L.P.C.). The method consists in performing on line detection of changes of the parameter Θ. The divergence test is based on the monitoring of a suitable statistic distance between two models Θ_1 and Θ_2. A change occurs when a threshold is exceeded. The procedure of detection is performed in parallel on the signal as on the high pass filtered signal. To avoid omissions, the signal is processed in the backward direction when the delay between two boundaries is too long (100ms). The parameters (AR order, thresholds) are speaker independent.

Voicing Test

A first test is applied to label segments as voiced/unvoiced/silence units. It is based on three parameters :

- the signal energy,
- the correlation of the signal
- the first reflection coefficient.

The result is adapted using the zero level crossing ratio. Each segment is then characterized as voiced or unvoiced. Voiced as unvoiced neighbouring segments are grouped together, and global frontiers give the VOW and VTW events.

Plosives Detection Test

Even if the automatic segmentation generally gives a rupture for the burst of the plosive, we have developped a complementary

centisecond test. It will permit a more robust detection. The test is based on two parameters :

- the formantic energy [500Hz-3000Hz],
- and the residual energy, under autoregressive model hypothesis.

The formantic energy permits to determine the potential vocal tract occlusion areas. A gaussian autoregressive model is calculated on each occlusion area on 2 ms sliding windows : a threshold crossing is equivalent to an abrupt opening of the vocal tract. We use adapted tresholds, function of the standard deviation over the occlusion areas.

The voicing test is applied on both sides of the detection : when a noise follows a silence, we detect the release of an unvoiced plosive SRW.

3.3. The Oral Signal

The recording technique for the aerodynamic signals is a pneumotach system using a Rothenberg mask. The drawback is the bad SNR of the signals and we first filter the signals with a classic low pass band filter.

As articulatory events square with changes of gradient, we perform a regression interpolation. The application of specific rules gives us the final labelling. The detection of CCO and CRO is difficult and specific : the closure CCO is linked to a high decrease of the oral airflow, during a more or less important delay. But the rules to detect these two events remain subjective and are still discussed with the experts.

3.4.The ElectroPalatoGraphic Patterns

As for the manual detection labelling, the automatic labelling is a dynamic process through the closure areas.

For closure, we define two masks according to the two different closure configurations : palatal or velar. We first detect the closure areas, that is to say the image sequence revealing an occlusion. The boundaries of the closure area precisely indicate the CCE and the CRE labels. The ACE is detected according to the place of the closure. It is a pattern in which there is a sufficient number of contacts around the center of the closure place. The MCE is the first pattern in the closure area, in which the number of contacts in the closure place is maximum. This method permits to determine the exact place of closure.

4 RESULTS AND DISCUSSION

4.1. Corpus

The corpus is composed of ten repetitions of ten sentences for each speaker, and of isolated words and logatomes. We have five different speakers male and female.

To assess AMULET, we measure the delay between the manual and the automatic articulatory events. As the dataset of manual labelled signals is restricted, our evaluation is done on five repetitions of two sentences pronounced by two speakers :

« La cousine de Vichy épousa un hippie à Toulouse »
« C'est maintenant que la smala les acclame ».

4.2. The acoustic signal

We have measured the delay between the automatic and the manual labelling under 10 ms, under 20 ms and over 20 ms. We have also took the omissions (O) and the insertions (I) into account. Results for the acoustic signals are presented in table 3.

	<10	10<<20	> 20	O	I
VOW	69/73	1/73	3/73		2
VTW	61/73	4/73	8/73		2
SRW	25/39	2/39	1/39	11/39	8

Table 3 : Number of automatic labels vs manual ones. Delay in ms.

Delays greater than 20 ms are often due to a persistent sinusoidal wave. Insertions of VOW and VTW will be interpreted in the future, with supplementary treatments, as consonantic areas.

Our automatic detection of plosives release detects all bursts of /t/ and /k/, as well as /p/ in a high frequency context. Most omissions of SRW are /p/ in context /u/.

4.3. The oral signal

Results under 20 ms are quite good (table 4). Manual labelling is often too fine, and criteria can vary with the experts. We can observe a deficience for the two events CCO and CRO, this is mainly due to a lack of objective criteria. This is discussed with the experts.

	<10	10<<20	> 20	O	I
BFO	49/93	16/93	9/93	19/93	5
DFO	64/101	23/101	4/101	10/101	6
MFO	67/83	12/83	1/83	3/83	40
mFO	19/53	16/53	3/53	21/53	2
CCO	17/32	4/32	2/32	9/32	3
CRO	27/41	3/41	5/41	6/41	1

Table 4: Number of automatic labels vs manual ones. Delay in ms.

4.4. The ElectroPalatoGraphic patterns

Detection of CCE and CRE is very robust (table 5). We precisely define the time of closure and release and the place (palatal or velar). Detection of the approach of closure depends on the preceding context, this explains delays greater than 20 ms. Criteria are discussed with the experts in order to define a new strategy of labelling for this event.

	< 10	10<<20	> 20
ACE	58/77		18/77
CCE	78/78		
MCE	64/72	4/72	4/72
CRE	78/78		

Table 5 : Number of automatic labels vs manual ones. Delay in ms.

The manual annotation for the event MCE can be the result of the application of different strategies. We have chosen the most frequent one, and the most interesting. This explains delays greater than 10 ms. Qualitative results are also very good . For example, we detect the overlapping movement in the double closure /kl/.

5. ACOUSTIC AND ARTICULATORY CORRELATIONS IN VOICELESS PLOSIVE PRODUCTION

The main goal of this study is to develop a robust automatic voiceless plosive detection and identification system.

On the acoustic signal, we can detect an articulatory event revealing a voiceless plosive release SRW. This event permits to detect a voiceless plosive. Nevertheless, we have three problems :

- the release is not ever realized, for example for /p/ and /u/,
- we have some omissions due to our system,
- we have also some insertions.

In order to have a more robust detection, we can study articulatory informations. Two questions arise : what are the interesting articulatory gestures ? and what are their spatio-temporal correlations ?

In order to answer these two questions, we have study the three plosives /p/, /t/ and /k/ in four repetitions of four French sentences for two speakers, male and female. This study is based on the automatic annotation.

Two articulatory gestures are used, the complete constriction on the oral signal, and the complete constriction on the E.P.G. (table 6).

		/i/	**/a/**	**/u/**
/p/	Oral	13/14		6/6
	EPG	0		0
/t/	Oral		7/7	17/18
	EPG		6/7	17/18
/k/	Oral		6/6	9/9
	EPG		5/6	9/9

Table 6 : Number of apparitions of the different articulatory gestures.

For /p/, the oral constriction is always present. /p/ is a labial plosive, so we never detect the presence of the E.P.G. constriction. We can observe that the articulatory event CCO corresponds to the beginning of the silence, and the articulatory event CRO corresponds to the SRW event with some milliseconds of delay after the SRW detection.

For /t/ and /k/, strategies are the same. Oral constriction as well as EPG constriction are very often present. Moreover, the place detection of articulation is always palatal for /t/ and velar for /k/. We can observe that the articulatory event CCO corresponds to the beginning of the silence, and the articulatory event CRO corresponds to the SRW event with some milliseconds of delay after the SRW detection. We can observe that the articulatory event CCE corresponds to the beginning of the silence or to the persistent sinusoidal wave after the end of voicing of the preceeding sound, and the articulatory event CRE corresponds exactly to the SRW event.

These results show that we can use the informations of different articulatory channels in order to give an accurate detection and identification of voiceless plosives.

6. CONCLUSION

As a conclusion, we first prove that the automatic labelling of multi-sensor speech data is feasable.The automatic labelling system developed already gives good results : about 80% of good detections under 10 ms except for CCO and CRO and about 90% for the E.P.G. . Some discrepancies are due to the systematic nature of our procedures, and others to the manual labelling criteria variations.

The adjusting of the automatic labelling has permited an interactive learning between the experts and the machine ; it has also permited to assess and to precise the manual labelling criteria.

Phoneticians are interested in these results for many reasons. First, the automatic labelling ensures the channel-independency of the annotations and it permits a robust application of defined criteria. The automatic procedure is also an important timesaver, it will permit to label large databases in order to perform statistics.

7. AKNOWLEDGEMENTS

The author wishes to thank Régine André-Obrecht for her efficient assistance and contributions as well as for her friendly support.

8. REFERENCES

1. J. VAISSIERE, " Speech recognition : a tutorial ", ed. F. Fallside and W. A. Woods, Prentice Hall International, pp 191-236, 1986.
2. A. M. LIBERMAN, I.G. MATTINGLY, " The motor Theory of Speech Perception Reversed ", *Cognition*, 21, pp 1-36.
3. A. MARCHAL, N. N'GUYAN-TRONG , " Non Linearity and phonetic Segmentation", J. Acoust. Soc. Am., Suppl.1, Vol87, pp79-82, 1990.
4. N. PARLANGEAU, R. ANDRE-OBRECHT, A. MARCHAL, " Automatic annotation of a multi-sensor speech database ", ICASSP 96.
5. C.P. BROWMAN, L. GOLDSTEIN, " Articulatory gestures as phonological units ", Phonology, 6, pp 201-251.
6. C.A. FOWLER, L.D. ROSENBLUM, " The Perception of Phonetic Gestures", Hashins L. Status Report on Speech Research,1989, SR-99/100,pp 102-117.
7. R. ANDRE-OBRECHT, " A new approach for the automatic segmentation of continuous speech signals", *IEEE Trans on ASSP*, Jan. 1988, Vol. 36 no 1, pp 29-40.
8. K. ERLER, L. DENG, " HMM representation of quantized articulatory features for recognition of highly confusible words", ICASSP 92, Vol. 1. pp 545-548