A DESCRIPTIVE FRAMEWORK FOR THE INVESTIGATION OF SPATIO-TEMPORAL RELATIONSHIPS AMONG TRACK VARIABLES

*Nathalie Parlangeau *- Régine André-Obrecht *- 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 in order to propose an intermediate level of representation and to assess gestural phonetic theory. We present here the framework of this investigation, the automatic labelling of the multi-sensor speech database ACCOR.

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 immediatly reflect the underlaying organization. The question that arises 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.

The first step of our study consists in automatically labelling the multi-sensor speech database which has been 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 trace and the ElectroPalatoGraphic patterns.

METHODOLOGICAL FRAME

Considering that speech is the output of a production process which relies for its execution on coordinated gestures, the annotation should reflect articulatory timing ; what must be located are articulatory events and not segments [3]. The annotation of the database is based on the following two principles :

- **non-linearity**,
- **channel-independency** of information.

The first principle is adopted to lead to proper annotation and to not 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 levels of representation. 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 indications of oncoming gestures from and towards articulatory goals. They are marked in the temporal domain according to precisely defined criteria.

THE ACOUSTIC SIGNAL

Articulatory goals

The labels currently used by phonetician experts, on the acoustic signal, are :

-VOW and VTW, Voice Onset and Voice Termination,

-SCW and SRW, Stop Closure and Stop Release of plosives /t/ and /k/.

For phoneticians, the label SCW means "a silence before a stop release" ; to preserve the non-linearity and to obtain systematic detections, we prefer to interpret this label as a simple closion before a silence.

Labelling methods

We first detect the acoustic discontinuities using a robust automatic segmentation method, the Forward-Backward divergence method [4] : 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 a 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 backwards when the delay between two boundaries is too long (100ms). The parameters (AR order, thresholds) are speaker independent.

Follows a first test to label segments as voiced/unvoiced/silence units. It is based on the mean variations of the energy, the correlation of the signal and the first reflection coefficient. The result is adapted using the zero level crossing ratio.

Next, we use a plosive detection test based on a Fourier Transform [5]. Two functions, the formantic energy Δ_n, and the high frequency energy variation Λ_n, are monitored. To detect a plosion, Δ_n must be lower than a threshold T1 and Λ_n must be higher than a threshold T2. We so locate voiced as well as unvoiced plosive bursts, the silence or the voiced segment before the burst.

Results and discussion

The events VOW and VTW are systematically found by our procedure, but we may observe a delay between the automatic position and the manual one. The table 1 gives an indication of these differences.

Large delays are specially present for the VTW event ; they are often due to a persistent sinusoidal wave which is present between the closure and the silence.

Table 1 : Number of automatic labels vs manual ones. Delay in ms.

	< 10	10<<20	>20
VOW	66/77	7/77	4/77
VTW	43/77	18/77	16/77

The SRW event corresponds to an unvoiced plosive burst. Our method detects the burst of all the phonemes /t/ and /k/, it detects also the labial plosive /p/ when it is located before anterior vowels.

THE LARYNGOGRAPH TRACE

Articulatory goals

Four articulatory events must be detected :

-VOX and VTX respectively Voice Onset and Voice Termination,

-PUX a Peack in an unvoiced segment,

-SGX a glottal stop.

Labelling methods

We use a simplified version of the Forward divergence method to detect the discontinuities of the laryngograph signal. Once the 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 ratio between the two level crossing rates. VOX and VTX events are finally labelled according to very simple rules.

On the unvoiced segment, the event PUX squares with a change of gradient, so we make a regression interpolation. The PUX label results of a temporal coordination between the regression variations and distance from the VTX and VOX labels.

Results and discussion

Table 2 : Number of automatic labels vs manual ones. Delay in ms.

	< 10	10 << 20
VOX	26/27	1/27
VTX	26/27	1/27
PUX	15/24	

Good results are obtained from the VOX and VTX labels. For the PUX event, nine events are not found. These results are due to the lack of manual precise criteria ; this point is discussed with the experts. We observe some insertions due to the systematic application of the PUX rules.

The SGX event is not automatically detected because we have a single realization on the french sentences.

THE AERODYNAMIC AIRFLOW TRACE

The aerodynamic signals are the nasal and the oral volume velocity traces.

Articulatory goals

The nasal events to be detected are :

- BFN and DFN respectively Build up of airflow and Decline of airflow.
- MFN Maximum airflow.

The oral events are :

-BFN and DFN respectively Build up of airflow and Decline of airlow,

-MFN and mFN respectively Maximum and minimum airflow,

-SCO and SRO respectively stop Closure and Release.

Labelling methods

The recording technique for these signals is a pneumotach system using a Rothenberg mask. The drawback is the bad SNR of the signals. It is a problem concerning an automatic labelling, so 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 for each signal.

Results and Discussion

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

	< 10	10<<20	> 20
BFN	42/65	7/65	8/65
MFN	3/12		
DFN	1/12	2/12	
MFDFN	43/49	1/49	2/49

We can see a seemingly bad result for DFN and MFN. In fact these two events are often labelled very closer, and our system detects an MFDFN event. Most omissions are explained by a too fine manual labelling.

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

	< 10
MFDFO	49/49
mFDFO	24/32
SCO	13/22
SRO	24/28

Major problems occur when we have to detect the SCO event : SCO are generally confused with the mFO event and the criteria to avoid these substitions remain subjective.

THE EPG PATTERNS

Articulatory goals

The phonetician experts search events like Closure and Constriction and want to detect :

- for Closure,
  - ACE approach to closure,
  - SCE stop closure
  - MCE maximum closure
  - SRE stop release.
- for Constriction,
  - ACE approach to constriction
  - MCE maximum constriction
  - CRE constriction release.

Even if some events have the same labels, their detection depends on the context Closure or Constriction.

Labelling methods

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

The patterns are 16*16 point images representing the tongue contact points.

For closure, we define three masks according to the three different closure configurations : a front , middle or back closure. The boundaries of the closure area precisely indicate the SCE and the SRE 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.

For constriction, the method is now in progress ; we use the same approach to locate the constriction areas.

Results and Discussion

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

	< 10	10<<20	< 20
ACE	58/77		18/77
SCE	78/78		
MCE	64/72	4/72	4/72
SRE	78/78		

The SCE and SRE detections are very robust. We almost indicate precisely the closure place.

The ACE label detection depends on the previous context. If it is a constriction, the detection has to be quite different. We do not take this difference into account, that explains delays greater than 20 ms.

To label MCE, different manual strategies are observed and we choose the frequent one. The delays are explained by this difference of strategies.

First experiments show correct detections of constriction areas.

CONCLUSION

We define an automatic labelling system for a multi-sensor speech database and quite good results are obtained. Discrepancies are due to the systematic nature of our procedures, and to the manual labelling criteria variations. This work permits to assess and to precise the manual labelling criteria.

Phoneticians are interested in these results for many reasons. First, the automatic labelling ensure the channel-independency of the annotations and it permits a robust application of defined criteria. The automatic procedure is also an important timesaver.

This work is the framework for the investigation of spatio-temporal correlations among track variables. It will permit to study an alternative to previous articulatory models for Automatic Speech Recognition [6].

REFERENCES

[1] J. VAISSIERE (1986)," Speech recognition : a tutorial ", ed. F. Fallside and W. A. Woods, Prentice Hall International, pp 191-236.

[2] A. M. LIBERMAN, I.G. MATTINGLY, " The motor Theory of Speech Perception Reversed ", *Cognition*, 21, pp 1-36.

[3] A. MARCHAL, W. J. HARDCASTLE (1993), " ACCOR : Instrumentation and database for cross-langage study of coarticulation", Langage and Speech, 2-3,pp 137-153,.

[4] A. MARCHAL, N. N'GUYAN-TRONG (1990), " Non Linearity and phonetic Segmentation", J. Acoust. Soc. Am., Suppl.1, Vol87, pp79-82.

[5] R. ANDRE-OBRECHT (1988), " A new approach for the automatic segmentation of continuous speech signals", *IEEE Trans on ASSP*.

[6] F. MALBOS, R. ANDRE-OBRECHT, M. BAUDRY (1994), " Comparaison de deux méthodes non paramétriques pour le détection des occlusives sourdes " , *XXèmes Journées d'Etudes sur la Parole*, pp175-180,.

[7] K. ERLER, L. DENG (1992), " HMM representation of quantized articulatory features for recognition of highly confusible words", *ICASSP 92*, Vol 1, pp 545-548.