THE RELEVANCE OF BASIC RESEARCH IN ARTICULATORY PHONETICS TO SPEECH TECHNOLOGY

A. Marchal*, W.J. Hardcastle**

*Institut de Phonétique, URA 261, CNRS, Aix en Provence, FRANCE;
**Speech Research Laboratory, University of Reading, UK.

ABSTRACT: For many applications in speech technology, decisive progress would result from the availability of an articulatory representation of speech utterances. While the acoustic mapping of the geometry of the human vocal tract during speech articulation is well understood today, the solution of the inverse problem, namely reconstructing the articulatory processes from the acoustic information, is still unsolved. The coarticulatory phenomena as a source of information have been almost entirely neglected. We will present in this paper the research action "ACCOR" ("Articulatory-Acoustic Correlations in Coarticulatory Processes: a cross-linguistic investigation") which has been recently launched under the EEC-funded ESPRIT II/BRA Program. It integrates investigation of the coarticulatory regularities themselves with research into new and improved ways of exploiting these regularities in deriving articulatory representations through the acoustic analysis of speech.

INTRODUCTION

Continuous speech is characterized by a great variability in the acoustic attributes of segments. They are highly sensitive to context and bear little resemblance to the intrinsic characteristics manifested when they are uttered in isolation.

This fact constitutes an unavoidable problem in speech technology. Most of the solutions tend in fact to ignore the source of variability, reduce it with statistical power or even look for some invariance without any understanding of the speech encoding process.

However, the variability can itself become a productive source of information if we learn to model the underlying physiological and linguistic constraints affecting the dynamics of articulation. The main source of systematic variability at the segmental level is undoubtedtly coarticulation.

1. RATIONALE FOR A CROSS-LANGUAGE ARTICULATORY-ACOUSTIC INVESTIGATION OF COARTICULATION

1.1 Scientific Paradigm

From a theoretical point of view, current models of speech production are inadequate in their attemps to account for connected speech processes such as coarticulation. The so-called "translation theories" (feature spreading models) explain allophonic variations by the cyclic execution of series of transformations, in which speech is considered as the temporal concatenation of states of the vocal tract resulting from phonological processes with coarticulation resulting from features to and from neighbouring segments.

However, as Fowler (1980) and Linell (1982) demonstrate, this theoretical approach is unproductive in nature and can only lead to an epistemological dead-end. Coarticulation cannot be circularly defined as the translation of "uncoarticulated" segments into coarticulated ones where "uncoarticulated" is never defined without reference to the concept it attempts to explain. Coarticulation could rather be viewed as a coproduction process (Marchal,1988) where the different dimensions of the articulatory process are executed and controlled by a few coordinative structures with motor equivalence as basic principle accounting for the great plasticity of speech movements.

The production of speech can thus be tentatively described as the functionnal operation of the pulmonic, phonatory and articulatory systems acting as coordinative structures.

A speaker creates variations in air pressure and airflow in the vocal tract by varying the positions and trajectories of the lips, jaw, tongue body, tongue tip, the velum and the opening of the glottis. An essential property of the production process consists of the dynamic organization of the speech gestures and their intrinsic timing. A segment cannot be described in terms of static features but should rather be seen as a dynamic gesture.
Following this, our aim consists of very carefully describing the coordination of the activity of the different groupings of articulators and the resulting acoustic output. In addition, we can predict that a given articulatory process may operate differently in two languages with different phonological systems and we need to investigate the interaction of the two sources of variation.
For this purpose, a cross-language approach is used in the ACCOR project involving 7 languages that differ in their phonological and rythmical structure -English, French, German, Gaelic, Italian, Catalan and Swedish. Biomechanical factors beyond the conscious control of the speaker which are related to the inherent characteristics of the speech production apparatus, factors such as the mass, inertia, elasticity of the speech organs, the mechanical linkages between them and the neuromuscular properties of the cranial nerve system, are expected to be universal.

1.2 Acoustic-articulatory relations

Extracting articulatory features and studying coarticulation phenomena from the speech waveform has a long tradition within acoustic and articulatory phonetics. In recent years, this field also gained importance for speech recognition. Experiments showed that while speech-specific knowledge can be incorporated into speech recognition systems (Zue, 1975) such systems, which try to capture the variabilities by applying common rules, are in fact less robust and successful than systems which model variabilities exclusively with powerful statistical methods, e.g. Hidden Markov Models (HMM). This mainly depends on the fact that it is very difficult to integrate articulatory or phonetic knowledge into current commonly used feature extraction and classification modules of speech recognition systems. HMM statistical methods work very well for speaker-dependant recognition. But for speaker-independant recognition with a large vocabulary on small machine, we believe that articulatory features are essential. The usual way to take into account coarticulation effects in speech recognition systems is to use larger recognition units, such as di-syllables or diphones, but again this is certainly too rudimentary, if improvements in large vocabulary speaker-independant recognition are to be achieved. In general, until now, too little consideration has been given to articulatory feature extraction for speech recognition.

1.3 Acoustic-Articulatory Model

The models for speech production currently used for speech recognition are based on the one-dimensional propagation of sound waves in very simplified geometry. The vocal tract is regarded as a linear filter modulating the glottis impluse. Some authors have, however, drawn attention to the features not considered by this approach. (Kaiser (1983) for example, emphasized the importance of flow phenomena, e.g. vortices, on the generation of speech and Thomas (1980) simulated air flow and sound waves in a two-dimentional time-variant geometry representing the vocal tract).

2. DATA ACQUISITION AND DATA PROCESSING

2.1 Articulatory Dimensions

To carry out an investigation on coarticulation it is necessary to examine the activities of the major physiological systems underlying speech production: the respiratory system (producing a flow of air), the laryngeal system (modifying the airflow by the valving mechanism of the vocal folds) and the complex of supraglottal structures in the mouth and nose such as the tongue, lips, jaw and soft palate, shaping the vocal tract into different resonating cavities. The speech production process consists of the dynamic coordination of these structures. Our aim in the ACCOR project is a careful description of such coordination and the resulting acoustic output. In addition, by examining the details of a given articulatory process in a number of different languages, we will be able to determine how such processes differ according to the different phonological systems, and thus will be in a position to investigate interactions between the two sources of variation.

In view of their importance in speech production, the following articulatory dimensions will be examined for each speaker:

-tongue body and tongue tip/blade trajectories
-lips and jaw movements
-velum (soft palate) activity
-variations in volume velocity of air in the voacl tract
-laryngeal activity.

2.2 Techniques

Most of the articulatory dimensions are not susceptible to direct observation. So specialized instrumental techniques have been developed for investigating them. However, not all techniques are suitables for the ACCOR project. To be suitable for recording articulatory activity in the project, an instrumental technique must satisfy certain requirements:

-it must be non-invasive and offer no potential danger or discomfort to the subject;

-it should not appreciably interfere with normal speech production;

-it must have a frequency response suitable for handling the fast rapidly-changing activities of speech organs such as the tongue tip (up to 220mm/sec during normal speech);

-it must provide measures of speech data wich are phonetically relevant;

-it must be reliable, with calculable measurement errors within acceptable limits;

In the first phase of the project, a detailed review of available literature on the subject of instrumental analyses of articulatory processes, particularly coarticulation, was carried out.

The review highlighted the advantages and limitations of the different techniques, and was used to determine their suitability for the project. In view of the above, we have decided to use the following instruments in the initial data acquisition stage: electropalatograph, pneumotach (for oral and nasal volume velocity), laryngograph (for voicing registration), Movetrack (for jaw and lips movements), together with a D.A.T. for recording of the acoustic signal.

2.3 Multichannel recording

2.3.1 Data acquisition system

We need to simultaneouly investigate the activity of different motor subsytems of speech, the resulting movement trajectories and the acoustic output. The synchronization of the articulatory and the acoustic data is crucial for this work. To this end, the Speech Research Laboratory at Reading has developed a multichannel data acquistion system based on an IBM AT which enables the simultaneous recording of the acoustic signal and up to 5 additional channels (Hardcastle, Jones, knight, Trudgeon & Calder, 1989). The hardware and software for a prototype system was carried out as part of a collaborative project between the Speech Research Laboratory at Reading and the IBM (UK) Scientific Center, Winchester. Various modifications have been carried out to the system to tailor it to the needs of the ACCOR project. A parallel system, PHYSIOLOGIA, is currently under development at Aix en Provence. This system will enable the simultaneous recording of the acoustic signal and up to 16 physiological parameters. At the present stage of development the analysis sofware is complete and the data acquisition interface nearing completion (Teston, Galindo & Marchal, 1990).

It is possible now to import files from the Reading-IBM system into the editing software of PHYSIOLOGIA. When the physiologia system is completed, it should be feasible for all centres in the ACCOR project to duplicate the necessary hardware for the system and thus to acquire and analyse their own multichannel data.

2.3.2 Data base

The collection of articulatory and acoustic data is in progress. In designing the corpus for such a data base, several language-specific constraints were taken into account. These included phonotaxis, sound inventories and prosodic patterns. It resulted in three sections of the corpus (1) structured VCV nonsense items, (2) real words matching the structure of (1) and sentences illustrating the main connected speech processes in the different languages. 10 repetitions of the whole corpus have been recorded for 35 speakers (5 speakers for 7 languages).It results in a total of 12500 Mbytes of multichannel data which is now being segmented, labelled and analyzed (Hardcastle & Marchal, 1990).

3. THE NOVELTY OF THE APPROACH

The key feature of the components of the ACCOR action can be summarized as follows:

With regard to coarticulation this research is one of the first attempts to adopt a comprehensive cross-language perspective on the phenomenon. The importance of this for elucidation of the linguistic and physiological constraints involved has been emphasized above. Astonishingly, even less work has compared coarticulatory behaviour in different motor subsystems in the same speaker. This work can easily be done within the proposed experimental framework and should throw significant light on the currently obscure but important question of wether different motor strategies are available to speakers and what factors determine the choice of strategy.

This research forms one of the first extended endeavours to exploit the systematic variability in the acoustic speech signal in order to derive an articulatory representation of spoken utterances. This will be accomplished by using top-down knowledge of coarticulatory regularities and the degrees of freedom of the articulatory apparatus.

The work will, in particular, explore the optimum way of combining this knowledge with acoustic features,such as formants, that can be related comparatively directly to articulatory categories.

By combining this new approach to analysis with the sophisticated statistical techniques already available in speech recognition, it will be possible to systematically evaluate the potential of the analyses for improving feature-extraction and classification stages of speech recognition systems.

CONCLUDING REMARKS

The key feature of the research action as a whole is its synthesis of speech technology and speech science. It forms one of the first attempts to unite on the one hand basic investigations of the articulatory process and on the other hand the exploitation of the kinematic, aerodynamic and acoustic knowledge gained thereby in derivation of articulatory representations from the acoustic signal and derivation of the sound-ware from vocal tract geometry.

The advantage of this approach is that each of the two main strands provides the testing ground of the analysis and modelling procedures used in the opposite strand. Thus these procedures will be subject to a healthy process of continual refinement. Where areas of inadequacy become evident, for example in analysis of lingual kinematics, it will always be possible to return immediately to the raw data in order to seek more revealing representations. Ways of extracting maximum benefit for concrete speech-recognition tasks from the knowledge gained will also be able to be pursued in this interactive fashion.

BIBLIOGRAPHY

Farnetani, E., (1990), *The V-C-V lingual coarticulation and its spatio-temporal domain*, in W.J. Hardcastle & A. Marchal (Eds) Speech Production and Speech Modelling,pp.93-130 (Kluwer: Dordrecht).

Fowler,C., (1980), *Coarticulation and theories of extrinsic timing*, J. of Phonetics, 8, 113-133.

Hardcastle, W.J., Jones, W., Knight, C., Trudgeon, A., and Calder, G., (1989), *Newdevelopments in Electropalatography: A sate of the art report*, Clinical Linguistics and Phonetics, 3,1-38.

Hardcastle, W.J., and Marchal, A., (1990), *EUR-ACCOR; A Multi-lingual Articulatory and Acoustic Database*, International Conference on Spoken Language Processing, Kobe.

Kaiser, J.F., (1983),*Some observations of vocal tract operation from a fluid point of view*, Conf. on Physiology and Biophysics of voice, Univ. of Iowa, Iowa City, May 4-7.

Linell, P., (1982), *The concept of phonological form and the activities of speech production and speech perception*, J. of Phonetics, 10, 37-72.

Marchal, A., (1988), *Coproduction: Evidence from EPG data*, Speech Comm., 7, 287-295.

Teston, B., Galindo, B., and Marchal (1990), *Design and Development of a work station for Speech Production Analysis,* Proc.Verba 90, Alcatel Face, Rome, 400-408.

Thomas, T.J., (1980), *A finite element model of fluid flow in the vocal tract*, Computer Speech and Language, 1, 131-151.

Zue, V., (1975), *The use of Speech knowledge in Automatic Speech Recognition*, IEEE, 23,1.

AKNOWLEDGMENT

Support for this research is provided by the Commission of the European Communities under the auspices of ESPRIT II Basic Research Action.
The principle investigators in the ACCOR group are:
A. Marchal, CNRS Aix en Provence; W. Hardcastle, Univ. of Reading;
E. Farnetani, CNR Padova; O. Schmidbauer, Siemens AG; F. Casacuberta, Univ. of Valencia;
O. Engstrand, Univ. of Stockholm; D. Recasens, Univ. of Barcelona;
A. Ni Chasaide, Univ. of Dublin; P. Hoole, Univ. of Munich.