LINKS BETWEEN SPEECH PRODUCTION AND ACOUSTICS: A SKETCH

Shinji Maeda

Département SIGNAL, CNRS URA-820
Ecole Nationale Supérieure des Télécommunications
46, rue Barrault, 75634 Paris Cedex 13, France

ABSTRACT

This paper reviews some of issues related to speech production. It is not intended to give an overview. Rather our attention is focused to particular issues related, directly or indirectly, to the topics dealt by these papers presented in this symposium. We discuss the specification of speech at different levels in the production chain and the mapping relations between those descriptions.

1. INTRODUCTION

It is our basic assumption that linguistic messages, segmental or suprasegmental, and extralinguistic messages, such as speaker's emotion, are coded into acoustic signals of speech through the phonatory and articulatory processes. These messages are decoded by listeners, establishing speech communication. Presumably, the linguistic messages are represented symbolically by phonological forms within linguistic structures of a particular language. One of the major research goals has always been to break the code, *i.e.*, to relate the abstract symbolic representation in the linguistic description of speech to observable physical signals. We cannot expect a simple mapping relation from the symbolic description to observable physical characteristics, articulatory or acoustic, of the speech however. When we speak, a series of processes at different levels, cognitive, physiological, mechanical, and acoustic, is involved in speech production. The process at each level is a complex phenomenon. Probably, the processes do not operate in a simple sequential manner. Rather they interact in complex ways through both forward and backward pathways. The influences or constraints exercised at the different levels shapes "real-time" characteristics of speech as well as the phonetic structures of speech. These observations give us a good reason for studying the description of speech at different levels of speech production and the mapping relations at different successive levels.

Primarily due to technical difficulty in the acquisition of articulatory data, however, the majority of experimental works has dealt with acoustic data. Working with acoustic signals alone was justified by saying that "we speak to be heard" and that the signals are the intermediate media between the speaker and the listener. Presumably, the acoustic signals contain all information specifying the linguistic description of utterances. The simplest hypothesis can be that the acoustic signals contain context independent invariant features related to phonological forms, despite widely variable aspects of articulatory gestures for realizing the same phoneme in different phonetic contexts. In certain cases, the reality might be more complicated than that. When the portion of signal corresponding to the reduced initial /ka/ is extracted out from /kaka/ for example, what we perceived is /kɛ/ [18]. Only when an onset portion of the second vowel is added to the signal, listeners perceive the vowel /a/, indicating that the context is necessary for the vowel to be perceived as the intended vowel /a/ instead of /ɛ/. It is necessary, therefore, to understand how the characteristics of acoustic signals vary as a function of the contexts. We have one more reason to study speech on the articulatory levels where the context effects actually occur. It is quite possible that speech perception can be better explained with reference to the description at a level other than acoustics [6]. Although the emphasis was put on the articulatory studies, this does not mean that the acoustic studies are considered as less important. Contrarily, the articulatory studies become more meaningful when they are carried out with the acoustics as the reference. What is really important is to understand how processes at different levels are linked together to shape speech sounds.

In this paper, we do not intend to provide a comprehensive overview of various issues on speech production (for such a review, see [2], for example). We shall focus our attention to the description of speech at different levels of speech production processes, and to the mapping relations between those descriptions.

2. PHONATION

Aspects of the sound generation inside the vocal tract, which is essentially an areodynamic phenomenon, are often put aside, when the segmental characteristics are at issue. The source-filter theory established by Fant [5] has a profound influence in our way of seeing the acoustic characteristics of speech signals. Segmental attributes are often related solely to the filter, *i.e.*, acoustic characteristics of the vocal tract which are the function of its geometrical configuration. An articulatory gesture in speech production is not only to create acoustic effects. It is also to create aerodynamic effects. In order to generate fricative sounds, the constriction is narrowed voluntarily so as to increase particle velocity high enough for the formation of turbulence. In order to produce stop sounds, the constriction has to be closed momentarily to increase the air pressure inside the cavity behind the closure for the generation of consonantal release. Probably, time-varying spectral characteristics of the source sounds contribute considerably to the identity of certain consonants. It is nearly impossible to isolate the spectral contribution of sources from the acoustic signals, however. We need to observe directly the aerodynamic phenomena inside the vocal tract, which is also nearly impossible at present. An alternative is to carry out such measurements on a mechanical model of the vocal tract [14] or to observe in detail the time-varying tract configuration and estimate the aerodynamics by calculations.

The paper presented by Drs. KIRITANI, HIROSE, and IMAGAWA describes a novel instrumentation, a high speed video recorder in connection with an endoscope or a fiberscope, to observe the vocal-fold vibration. The instrument is already useful for clinical applications. Probably it is possible to derive quantitative data, for example time-varying glottal area, from captured video images. In that case, the perspective is an exciting one for speech research. Such data can be used to test theoretical works, for example the classical two-mass model of the vocal-folds [1]. In the past, suprasegmental aspects of speech were studied primarily with fundamental frequency variation along utterances. With glottal area data, we may have an access to evaluate the time variation of glottal pulse shapes, and thus of glottal spectra, which might contribute to signal lexical stress, tones, emphasis, and emotion. Such instrument would be very useful also for observing fast labial release gestures. The flow variation just after the release, which is critical for the spectral characteristics of the burst onset, can be estimated from the observed time variation of the lip opening area [8].

3. TRACT LEVEL

At frequencies below 3 kHz, the sound waves propagate only at the directions of the tract length. The cross-sectional area varying along its length, *i.e.*, an area function, therefore, characterizes the acoustics. The vocal tract is represented by a piece of a straight tube with one end closed (corresponding to the glottis) and the other end opened (to the lips). An arbitrary tube shape, i.e., cross-sectional area variation along its length, must be constrained to account for geometrical shapes of the human vocal tract. Often the shape is parameterized including, in particular, the constriction locations and their degree, that excise a strong influence on the spectral shapes [*e.g.*, 5, 15]. The parameterization of the tract shapes allowed researchers to investigate the mapping relation from the tract shape to the acoustics.

As early as 1955, Fant has studied the relationships between tract shapes and the acoustics in terms of formant patterns, now classical "nomograms", using a four-tube tract model [5]. Stevens noticed that the formant sensitivity to a small variation in the constriction location, specified by the distance from the glottis to constriction point, is not uniform along the vocal tract. Rather there are locations where the sensitivity is relatively high (formant patterns are unstable against a small variation in the constriction location), and other locations where the sensitivity is low, (formants are relatively stable). This observation has lead Stevens to the proposition of "quantal nature of speech" [16]. The concept is generalized to "quantal theory" [17]. Although the theory is still under controversy (*e.g.*, [11]), it seems to explain well the formation of speech sounds in a number of cases.

In their paper, Drs. MRAYATI and CARRE propose another type of model of the vocal tract area function. Instead of constructing a tube model to characterize the tract shapes, as described above, the tube is divided into specific "regions" having fixed lengths solely on acoustic consideration. The division into regions is strictly based on the formant sensitivities to the localized variation of the cross-sectional area. Region lengths remain fixed and only region areas are varied to specify a vocal tract configuration. This model effectively exploits physical properties, manifested on the sensitivity functions, which ensures the maximum modulation of formant frequencies with respect to the neutral uniform tube configuration. One might question how accurately such model can describe observed static or time-varying vocal tract shapes. Even though the model is capable of producing any observed temporal pattern of the formant frequencies, with appropriate variations of the region areas, it could be still merely an equivalent representation of the original formant variations. Nevertheless, it is quite appealing to ask whether or not humans also exploit such physical properties in speech production. If that is the case, they influence upon the formation of sound pattern of speech.

The specification of speech in terms of the tract configuration plays an important function in formulating a scheme for the articulatory control as described later. It serves for the specification of articulation goals. The search of articulatory correlates, for example, of vowel height, such as mandibular height or tongue dorsum height, constantly fails in the past, due to a large observed variabilities of these positions. Contrarily, the acoustic specification of vowels, typically with F1, turned out be much more consistent with the phonological notion of height. Wood has demonstrated, however, that the tract configuration, essentially in terms of the constriction location and the degree of the constriction, results in more coherent relation with vowel height [19]. This is not so surprise in the sense that the area function and the acoustic are tightly related by laws of physics, *i.e.*, the tract configuration determines the acoustic. (The inverse is not necessarily true.) It is reasonable, therefore, to set up an articulatory goal in terms of the tract shape,

which is equivalent to the acoustic goal. This equivalence, however, is meaningful only in the specification of the static configurations, such as articulatory goals.

In the description of the dynamic aspects of speech, the specifications at the tract level and at the acoustic level become distinctively different. The articulatory process involves both spatial and temporal organization. For example, a vowel-consonant-vowel (VCV) sequence is produced by a global vowel articulation of the tongue dorsum plus spatially more localized consonantal gestures with participation of the lips, tongue tip *etc.* [12]. For a longer sequence, such as VCVC.., consonantal gestures can be characterized explicitly in terms of the places and timing relations of the sequential consonantal gestures. The multidimensional organization of speech would appear on the single dimensional temporal pattern of the corresponding acoustic signals. One of difficulties in describing context dependent variability of the acoustic characteristics, such as coarticulation, is due to the fact that the specification of speech at articulatory level, which is inherently multidimensional, is mapped or linked to the single dimensional acoustic space.

Although the spatiotemporal organization of speech might be described at the tract level, there exists an inherent limitation in its effectiveness. In reality, the tract shapes are determined by the states of the individual articulatory organs. The spatiotemporal organization means actually temporal patterns and their inter-timing (phasing) relations of the individual articulators. In the cases where successive articulatory goals are the spatially localized and, say, anatomically separated, for example a sequence of gestures involving the labial, velic, tongue tip, or laryngeal manoeuvres, the description of the spatiotemporal organization at the tract level might correspond in a simple way to that at the articulatory level. In other cases where gestures simultaneously involves the activation of multiple of the articulatory organs, such as a sequence of vowels and consonants including velars, the correspondence can become complex, say, tangled up. Strictly speaking, even localized labial gesture involves the participation of not only the intrinsic labial manoeuvres but also the mandible movements. Therefore, even when the organization appears to be simple and clean at the articulatory level, the same organization can appear to be quite messy at the tract level, and probably worse at the acoustic level. These arguments direct us towards the specification of speech at the articulatory level, especially when the spatiotemporal organization is in question.

4. ARTICULATORY LEVEL

The specification at the articulatory level is characterized by its relatively large variability, as already mentioned before. If the observed variability is random and unexplainable, the articulatory description of speech would be of little value. The question might be raised is what factors cause such variability. One of the causes is motor equivalence. It has been evidenced that the variability of the composite product is significantly smaller than that of individual articulatory movements. For example, the variability in the lip aperture is smaller than that of the lip or of the jaw position [4]. From "bite-block" experiments, Lindblom, Lubker, and Gay concluded that speech production is compensatory in nature [7]. A deviation of the mandibular position, due to an external as the bite block or an internal organizational cause, can be compensated by the readjustment of the other articulators to restore the acoustic characteristics of speech. It may be noted that the motor equivalence is a mapping characteristic between the articulatory and tract levels, whereas the "compensation" is that between the articulatory and acoustic levels.

The paper presented by Dr. WOOD in this symposium adds more evidence to the compensatory phenomenon. From a model experiment, he has demonstrated almost perfect acoustic (Fl-F2) restoration in the mandible-labial coordination for the rounded vowels, such as /u/ and /o/. A sound change can occur between two proximate vowels, such as /u - o/, as the consequence of a single gesture, for example, the jaw opening or closing movement. Such sound change can be prevented, if necessary, by maneuvering the other articulators. We have shown also, in a similar model experiment, that such acoustic compensation occurs in the mandible-tongue dorsum coordination for unrounded vowels [9]. Moreover, at least for a limited case, such compensatory relation in the articulatory space can be specified by a simple proportional (or linear) relation between the positions of the paired articulators [10]. This implies that the mapping between the articulatory level and the acoustic level, by-passing the tract level, can be described in simple manner and that the acoustic goals can be specified directly at the articulatory level as relative potions of the specific paired articulators.

In order to model the observed spatio-temporal coordination, we must assume a function that controls the different articulators to achieve a series of targets, and that operates at levels higher than those for controlling the movements of the individual organs. Such a function is called "motor programming". Unfortunately, there is no means to observe directly the motor programming in operation. Electromyographic measurements are at present the best we can do for observing physiological patterns at levels higher than the articulatory level. Still they are limited only to the observation of individual muscle activities. It is important, therefore, to accumulate experimental data at the lower levels and to characterize their behaviors as precise as possible, in order to formulate the most plausible scheme for the articulatory control. The formulation depends on what we observed at lower levels. As mentioned earlier, the specification of speech at the tract level has a tight relation with that at acoustic level and exhibits relatively small variability. Thus in control scheme, such as a simulated feedback [7] or a task dynamic [13], the articulatory goals are specified at the tract level. It is also foreseeable to postulate another scheme, at least for vowels, in which the goals are specified directly at articulatory level, in terms of not absolute positions but of relative position among the individual articulators. In any case, any control scheme must take into account for such links among articulatory, tract, and acoustic levels.

Compensation is, probably, not only the factor involving the articulatory organization. The paper presented by Drs. ABRY and LALLOUACHE deals with the anticipatory articulation in the lip rounding gesture. The individual articulators tend to anticipate for production of upcoming string of segments, unless the anticipatory movement causes a sound change. Henke has explicitly implemented anticipation into the control scheme in his dynamic articulatory model, as "a lookahead operator" [3]. Alternative models for anticipation, as "time-locked", "hybrid", were proposed to explain observed data. The two authors demonstrated that the rounding anticipation could not be predicted by none of the three models and suggested that unpredictable data were due to prosodic effects which were not controlled in their data acquisition separated by two sessions. If this is the case, the segmental patterns in articulation are also influenced by the suprasegmantal factors, such as accents, grouping of words, *etc*.. This implies that the motor programming has to handle both segmental and suprasegmental requirement to issue appropriate commands to the individual articulators. If anticipation depends target positions specified for individual articulators, which also depend how compensation is employed, then intricate calculations involving both anticipation and compensation are required for the motor programming function. There is reason to believe, therefore, that the articulatory organization and thus speech production process is indeed a complex phenomenon.

REFERENCES

[1] Flanagan, J.L., Ishizaka, K., and Shipley, K.L. (1975). Synthesis of speech from a dynamic model of the vocal cords and vocal tract. *The Bell System Technical Journal,* **54**(3) 475 - 506.

[2] Fujimura, O. (1990). Methods and goals of speech production research. *Language and Speech,* 33(3), 195 - 258.

[3] Henke, W.L. (1966). *Dynamic Articulatory Model of Speech Production Using Computer Simulation.* PhD thesis, Department of Electrical Engineering, MIT.

[4] Hughes, O.M. and Abbs, J.H. (1976). Labio-mandibular coordination in the production of speech: Implication for the operation of motor equivalence. *Phonetica,* **33**, 199 - 221.

[5] Fant, G. (1960). *Acoustic theory of speech production.* The Hague: Mouton.

[6] Liberman, A.M. and Mattingly, I.G. (1985). The motor theory of speech perception revised. *Haskins Laboratories: Status Report on Speech Research,* **SR-82/83**, 63 - 93.

[7] Lindblom, B., Lubker, J., and Gay, T. (1979). Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. *Journal of Phonetics,* **7**, 147-161.

[8] Maeda, S. (1987). On the generation of sound in stop consonants. *Speech Communication Group Working papers,* Research Laboratory of Electronics, MIT, **5**, 1 - 14.

[9] Maeda, S. (1990). Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In W.J. Hardcastle & A. Marchal, (eds.), *Speech Production and Speech Modeling,* pp. 131-149. Kluwer Academic Publishers.

[10] Maeda, S. (1991). On articulatory and acoustic variabilities. (to be appeared on *Journal of Phonetics.*)

[11] Ohala, J.J. (ed.), Theme issue on the quantal nature of speech. *Journal of Phonetics,* **17** (1989).

[12] Öhman, S.E.G. (1967). Numerical model of coarticulation. *J. Acoust. Soc. Am.,* **41**, 310 - 320.

[13] Saltzman, E.L. and Munhall, K.G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology,* **1**(4), 333 - 382.

[14] Shadle, C. (1985). *The acoustics of fricative consonants.* PhD thesis, Department of Electrical Engineering and Computer Science, MIT.

[15] Stevens, K.N. and House, A.H. (1955). Development of a Quantitative Description of Vowel Articulation. *J. Acoust. Soc. Am.,* **27**(3), 484 - 493.

[16] Stevens, K.N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In E.E. Devid and P.B. Denes (eds.), *Human Communication: A Unified View,* 51 - 66. New York: McGraw Hill.

[17] Stevens, K.N. (1989). On the quantal nature of speech. *Journal of Phonetics,* **17**, 3 - 45.

[18] Vaissière, J. (1987). Effect of phonetic context and timing of the F-pattern of the vowels in continuous speech. *Proceedings of the 11th Congress of Phonetic Sciences,* Tallin, Estonie, 43 - 46.

[19] Wood, S. (1979). A radiographic analysis of constriction locations for vowels. *Journal of Phonetics,* **7**, 25 - 43.