The Production of Speech

Edited by
Peter F. MacNeilage

With 90 Figures

Springer-Verlag
New York Heidelberg Berlin

Peter F. MacNeilage
Departments of Linguistics and Psychology
The University of Texas at Austin
Austin, Texas 78712, U.S.A.

Library of Congress Cataloging in Publication Data
Main entry under title:
The Production of speech.
Papers presented at a conference on the production of speech, held at the University of Texas at Austin, on April 28–30, 1981, sponsored by the Center for Cognitive Science, and other bodies.
Bibliography: p.
Includes index.
1. Speech—Physiological aspects—Congresses.
I. MacNeilage, Peter F. II. University of Texas at Austin. Center for Cognitive Science.
QP306.P693 1983 612'.78 82-19246

Typeset by Ampersand, Rutland, Vermont
Printed and bound by Halliday Lithograph, West Hanover, Massachusetts
Printed in the United States of America

9 8 7 6 5 4 3 2 1

ISBN 0-387-90735-1 Springer-Verlag New York Heidelberg Berlin
ISBN 3-540-90735-1 Springer-Verlag Berlin Heidelberg New York

Preface

This monograph arose from a conference on the Production of Speech held at the University of Texas at Austin on April 28–30, 1981. It was sponsored by the Center for Cognitive Science, the College of Liberal Arts, and the Linguistics and Psychology Departments. The conference was the second in a series of conferences on human experimental psychology: the first, held to commemorate the 50th anniversary of the founding of the Psychology Department, resulted in publication of the monograph *Neural Mechanisms in Behavior,* D. McFadden (Ed.), Springer-Verlag, 1980.

The choice of the particular topic of the second conference was motivated by the belief that the state of knowledge of speech production had recently reached a critical mass, and that a good deal was to be gained from bringing together the foremost researchers in this field. The benefits were the opportunity for the participants to compare notes on their common problems, the publication of a monograph giving a comprehensive state-of-the-art picture of this research area, and the provision of enormous intellectual stimulus for local students of this topic.

The conference also provided an opportunity to honor Dr. Franklin Cooper, former President of Haskins Laboratories, who delivered the opening address at the conference, for his important research contributions to this area, his influence in fostering development of the area and, for want of a better phrase, his monumental good-citizenship. This purpose of the conference gave the present author (P.F.M.) particular pleasure, as the six-year period he had spent on the staff at Haskins Laboratories was the major formative influence on his thinking about speech production.

Chapter 12

In Favor of Some Uncommon Approaches to the Study of Speech

M. Y. LIBERMAN

I. Introduction—Phonology vs. Phonetics

It is no secret that phonetics and phonology are two very different cultures, despite the close logical connection between their nominal aims. In education, in terminology, in research methods, in styles of argument, and in scientific journal and professional society allegiance, the divergence is striking.

When members of one group choose to take notice of work in the other camp, an undercurrent of something akin to ethnic prejudice is commonly quite plain. Most phonologists (at least unconsciously) subscribe to Trubetzkoy's dictum that phonetics is to phonology as numismatics is to economics. I do not know of an equally celebrated phrase to express the contrary prejudice; the phonetician's rejoinder might be that phonetics is to phonology as physics is to theology, except that theology is of insufficiently low academic status for this to have quite enough sting.

The editor of this volume has suggested that I play the role of phonologist among the phonetic lions. Such a role does not fit perfectly, but it offers me the opportunity to discuss this cultural divergence, which I think is becoming more and more harmful to work in both fields, and to describe some issues (in linguistics generally, and in speech production research in particular) where there is real promise for progress based on an approach that does not fit easily into the traditions of either group.

The notion of phonological analysis, in one form or another, underlies almost all work on human speech communication. This idea, in its basic form, holds that the sonic identity of the words of a given language can be expressed by combining elements drawn from a small set of semantically meaningless primitives. These primitives may be in the nature of phonemes, features, morae, or whatever, and the methods of combination range from simple concatenation to deployment in complex syllabic or even suprasyllabic structures.

In the two and a half millennia since Panini, in a dozen or so traditions around the world, those who have undertaken to study human speech and language have generally come to some form of this conclusion. Furthermore, each tradition of analysis seems to arrive fairly quickly at a particular hypothesis about the phonetic decomposition of the language or languages of local interest, a hypothesis that has what we might call "transcriptional adequacy." In other words, each such system of description offers a way to characterize the pronunciation of each word in the language, and perhaps goes on to differentiate some variant pronunciations, if this is desired. (In the case of Panini, it is at least etymologically misleading to call this representation a "transcription," since the Paninian tradition was apparently an oral one.) The discovery and codification of such a transcription system is always a big step forward, entirely apart from any practical use as an orthography, since it permits the collection of observations about spoken language and provides a common terminology for the discussion and elucidation of their meaning.

Once discovered, such descriptive systems are easily accepted and commonly used in the systematic study of speech; their fate as orthographic methods is more erratic, depending on cultural circumstances of a largely irrational nature. The success and usefulness of these descriptive systems cannot be solely attributed to the penchant of human rationality for analytic decomposition. There is an equally long tradition of attempts to reduce lexical semantics to the combination of elements drawn from a small set of primitives, and the results have never been very useful or very generally accepted.

The modern tradition of research in spoken language, of which we are all a part, assumes (informally) a system of description that reached essentially its current form in the work of Daniel Jones, for the case of English, and has been variously extended and modified to cover other languages. The originators of this descriptive system assumed that its categories were physically definable. To date, no attempts to redeem this assumption by construction of explicit algorithms (operating on either acoustic or articulatory measurements) have been successful. This failure to give objective status to such representations has caused a certain general nervousness, but has done little to change their usage in practice. If we leave aside the question of what such transcriptions really mean (i.e., how to reconstruct them theoretically), and if we permit typographical detail to count for nought, then there is an astonishing uniformity of practice in casual notation among researchers who otherwise agree on nothing. Almost no one accepts this style of transcription as a theoretical construct, but almost everyone uses it as a notational convenience.

Indeed, it is really much more than a convenience—it is to a large extent the unconscious object of our discourse, our initial method of organizing and characterizing the phenomena we intend to study. It is important to stress the pretheoretical and informal nature of this common description—any attempt to fix its details, or determine its exact status, would soon dissolve in a welter of recriminations. Nevertheless, we all rely on our own comfortable versions of this way of speaking about speech.

There are many approaches to phonetics, and many approaches to phonology, and the number of descriptions and justifications of the division of labor between the two fields is probably not much smaller than the Cartesian product of the two sets of schools. The general acceptance of the division between fields, and the cultural divergence that ensues, cannot easily be explained by reference to any particular instance of such theories. Instead, the foundation of the split is to be found in our common acceptance of the pretheoretical system of descriptive categories just discussed. Since everyone implicitly accepts this way of talking about the stuff of speech, it is perfectly natural for one group of researchers to concentrate on the relation of such descriptive categories to the physical realities of articulation and sound, while another group tries to understand the intricate patterning of the corresponding pretheoretical categories under morphological inflection and derivation, optional pronunciation, and so forth.

Most of the differences in style and method between the two fields then follow, as natural adaptations of the organism to its research environment. I will return to this point later, and argue that many of these divergences have become seriously counterproductive, at least in the study of some important questions in both fields.

To avoid misunderstanding, I hasten to add that I do not question the appropriateness of any division of labor in research on speech and language. Surely the study of Semitic morphology and the modeling of oronasal coupling require rather different knowledge and skills, and it would be unreasonable to insist that no one could undertake one of these who could not also handle the other.

II. Intonation: Where a Hybrid Approach Is Required

Much of my own research has dealt with English intonation. In this area, the comforting common ground of pretheoretical description is conspicuous by its absence. There is no generally accepted practice for notating, naming, or even categorizing intonations. A number of schools of description exist, but their systems are usually difficult for outsiders to learn, the completeness of their coverage is often suspect, the intersubjective reliability of their descriptions is

generally dubious, and the basic categories employed have (at least superficially) very little similarity from school to school. The prospect facing a newcomer to the field is indeed depressing, unless he or she has a perverse taste for intellectual balkanization.

The situation improves slightly with time. Experience with the phenomena of intonation, and time spent poring over the examples in the literature, lead eventually to the suspicion that the various authorities are indeed talking about the same things in different ways, roughly in the sense that romaji and hiragana transcriptions both characterize the same Japanese language. But this suspicion does not, at least for me, develop into a comfortable pretheoretical description.

It is not clear whether these difficulties reflect a basic difference between intonation and "segmental" phonology. I suspect that they do. Some obvious symptoms of the difference are intonation's lack of referential meaning; its deficiency in the equivalent of "word constancy" (by which I mean the homely conviction that "rabbit" cooed and "rabbit" snarled are instances of the same word); and its preoccupation with apparently gradient distinctions.

I will not speculate here about the cause of these symptoms. All that matters to the present discussion is their effect, which is to make the usual phonetics–phonology distinction inapplicable. Regardless of background, all students of intonation must think for themselves about what the basic categorization of intonational phenomena should be before they can begin even an informal investigation. Their research is (or should be) constantly drawn back to this fundamental question: Each advance in the basic categories of description permits the interpretation of a broader range of data, which often suggests a new modification of the initial descriptive assumptions.

In the face of such a complex problem, it is natural to use the widest available set of methods; these tend to complement one another, allowing a clearer fix on the object of study than an approach from only one side would. In trying to puzzle out the nature of intonational patterns, I and various colleagues (principally Alan Prince, with whom I have worked on theories of stress, and Janet Pierrehumbert, who is responsible for much of the tonal theory sketched below) have used at least the following list of methods:

1. Reliance on linguistic theories of stress and phrasing, insofar as they are helpful, and revision of such theories where they are not;
2. Impressionistic categorization of intonations, based on progressive ear training, coupled with informal analysis of instrumental F_0 contours;
3. Construction of explicit models for the abstract representation of intonation, in which we try to be clear about what the primitive concepts are, how they can be combined, how they are related to other aspects of speech and language, and how they are to be used in accounting for our observations;
4. Perceptual experiments to test the distinctness of intonational types, their appropriateness to various contexts of utterance, and the dependence of subjective pitch height on phrasal position;
5. Testing of phonological descriptions of intonation by synthesis: Explicit rules are used to construct F_0 contours from a "phonemic" representation, and the result is recombined with spectral parameters taken from a natural utterance, for evaluation by listening;
6. Modeling of the effects of length, order, and pitch-range variation on F_0 contours of the "same" intonational type.

The first three methods are typical of phonologists' practice; the last three methods are very much the sort of thing that phoneticians and psycholinguists do. I believe that we have made quite a bit of progress by using this combined approach; readers curious about details are referred to the bibliography. Our work would have had very little success, in my opinion, if we had taken either a purely phonological or a purely phonetic approach. Nor could we have succeeded by relying on the literature for either our phonological or our phonetic ideas—our results have depended on the interaction of new research in both areas.

Our conclusions have changed somewhat as the work has proceeded, and I am not yet confident that we have got the basic framework of description right. However, I have come to regard this feeling of uncertainty as a good thing, and would move on to new problems if it ever went away.

Of course, constant reevaluation of first principles in search of better theories is the normal method of science, and has been characteristic of most good work on language and speech. It is much less common to use a wide variety of methods, drawn from the traditions of both phonology and phonetics, in a concerted effort to clarify some aspect of human speech, considered simultaneously as a system of signs and as a signaling procedure.

The Hybrid Nature of the Resulting Theory

I have argued that some special properties of the field of intonation research make it appropriate to use hybrid methods in arriving at a theory. But it is interesting to note that the result is, in a sense, a hybrid theory. It is not that levels of analysis are inappropriate here—on the contrary, the theory proposes a rather clean and simple level of phonological analysis (involving sequences of high and low tones aligned with the stress pattern of the phrase), a few gradient "paralinguistic" dimensions (e.g., pitch range), and a small set of explicit principles for phonetic realization, which together are asserted to assign a (usually unique) description and derivation to every possible F_0 contour in the language. The division of the analysis into components is fairly traditional. However, the resulting theory can make sense of (even simple) examples only when all three aspects are brought into play at once.

The "phonetic realization principles" are very limited in power and scope—they work from left to right through the utterance, essentially one stress group at a time. However, they depend on local phonological structure, their local output usually depends on the pitch level assigned to a previous tone, and a certain

amount of interpolation occurs across tonally unspecified material, so that the correspondence between tones and local F_0 levels or configurations becomes somewhat opaque. Therefore, the phonological analysis does not offer a convincing account of an F_0 contour without the help of the phonetic implementation principles, together with assignment of values to gradient parameters such as relative prominence and pitch range.

It also appears that the "phonetic implementation principles" for intonation differ somewhat from language to language, although too few languages are well described for this point to be compelling to a skeptic.

III. The Relevance of a Hybrid Approach to Segmental Problems

Most of the arguments for a hybrid approach to intonation research also apply to the traditional problems of segmental phonetics and phonology; the need is just a little less obvious. Authorities in the field disagree about the fundamental characterization of the phenomena to be explained, and there is always a value in trying to work toward the object of study from several directions at once. Furthermore, there is good reason to suppose that much of the traditional data in segmental phonetics and phonology requires an essentially hybrid explanation, in the sense previously suggested for the case of intonational data.

A. The Case of Phonology

Within and among the various schools of generative phonology, there is increasing uncertainty about the basic nature of phonological representation. This trend can best be understood in a historical perspective. In the beginning, generative phonology relied largely on data taken from its predecessors; in the Sound Pattern of English (SPE), the contents of Kenyon–Knott and of Trager–Smith were reanalyzed and explained in terms of feature theory, ordered values, the cycle, and so forth. An explicit analogy to the Copernican reanalysis of Ptolemaic data was noted. The output of the SPE phonology could be translated into the notation of Trager–Smith by simple substitution of symbols; the new theory explained the old data by direct generation. The main representational innovation was feature theory, which remains easily intertranslatable with segmental alphabet theories. The rest of the new explanation depended on innovations in the rule systems.

Over the last few years, a number of new representational devices have been introduced into the generative armamentarium, principally the constructs of so-called autosegmental and metrical theories. Various representations of syllabic and suprasyllabic structures have been suggested, and features are assumed to migrate about in a variety of trees and graphs defined as the need arises. Phonological rule systems can be simplified, almost trivialized, at the cost of such enrichment of the structures they modify, and typological properties of phenomena such as stress, vowel harmony, and vowel epenthesis are elucidated.

I am basically in favor of such approaches to phonology, but it must be admitted that they bring a new kind of uncertainty with them. In the context of SPE-era phonology, one end of the problem was essentially fixed. Everyone agreed about what the system was supposed to generate (aside from arguments about how to spell out phonetic symbols in featural terms); the only argument was about the nature of underlying representations and the intervening rules. Now that arbitrary new representational structures are up for consideration, the desired form of the system's output is much less well defined. The logic of the situation has not really changed at all, but as a practical matter, the uncertainty about representations is much greater than it was, and there is a greater need for phonetic or psychological evidence to help constrain the choices.

Few phonologists would argue about the benefits of converging evidence from other sources, but the next point will be more controversial. There is an increasing amount of evidence, I think, that much of the traditional domain of phonological data actually belongs to a component whose function is analogous to that of our intonational implementation principles. Specifically, phonologically transparent (not lexically governed) allophonic variation seems to belong with a larger class of phonetic regularities that are not well modeled as feature- or structure-changing rules. Such regularities are usually dependent on phonological environment, not just on the superficial physics of surrounding articulations, but their consequences are gradient, apparently linked to the inherent dimensions of articulation, and modulated by prosodic and paralinguistic parameters. These implementational regularities have language-particular and indeed dialect-particular aspects; this, along with their dependence on phonological environment, makes it seem unlikely that they can entirely be explained by reference to the physics and physiology of the vocal organs. If I am right about the characterization of these regularities, then it is plausible to suppose that they represent the higher level aspects of speech motor control.

This point of view raises serious questions about the types of regularities that ought to be expressed by manipulation of phonological representations. This is a more serious form of uncertainty than the one mentioned earlier, and it requires (rather than simply invites) investigation by hybrid methods. Out of historical necessity, phonology has assumed (at least in practice) that its task was to explain the patterning of information in a class of well-defined symbolic objects, namely, phonetic representations. This convenient fiction has become more and more counterproductive, and should be gradually abandoned, as opportunity permits.

B. The Case of Phonetics

The interpretation of segmentally related acoustic and articulatory measurements has a lot in common with the interpretation of F_0 contours; the main difference, as mentioned earlier, is that there is a generally accepted informal classification of segmental categories available to rationalize the task. Aside from this initial advantage, very similar problems arise. There are obvious effects of phonological environment, of prosodic and paralinguistic variation, and of physical coarticulation that collectively make the connection between phonetic category and physical measurement anything but transparent. There is no choice of unit, even up to the word or the phrase, that entirely avoids this problem. In particular, the consequences of rate, emphasis, and style of speech are complex and pervasive. Such variation prevents even phrase-level units for a single speaker from having a straightforward physical definition.

Furthermore, if I am right about the nature of the implementational regularities mentioned earlier, better theories of the physics and low-level physiology of speech, although obviously desirable, will by no means suffice to explain the complex relation between words and sounds. Such an explanation requires explicit modeling of the nature of an utterance plan and of the process by which it is spoken. If the phenomena of allophonic variation are (even in part) consequences of the realization process, then the utterance plan must be rather more abstract than the standard forms of phonetic representation would suggest, and must be sufficiently rich in structure to condition the relevant regularities. To have any value as predictors of real data, the realization model must allow for the effects of local environment (in the plan) on the units that are manipulated, whatever these are to be, and for the effects of stress pattern, phrasing, rate, and so forth on the realization of the plan as a whole. Obviously, the physics and low-level physiology of the vocal organs must be employed to explain what they can.

I do not believe that any observationally adequate models of this kind now exist. Speech synthesis systems are the closest overall approximation, but their treatment of the consequences of contextual, prosodic, and paralinguistic variation is a rather erratic fit to measurements of natural speech, in my experience, and their internal workings are generally determined more by engineering expediency than by any consideration of theory. The construction of a complete model in an entirely principled way will presumably not be possible for a long time. Many partial successes are possible in the interim, but I strongly suspect that progress depends on an approach that gives serious thought to representational issues, while using these representations in explicit modeling of appropriate measurements.

Historically, phonetics has generally assumed that its task is to explain the physical realization of a class of pretheoretically defined categories, which are essentially those implied by traditional phonetic representations. Phonetics' physicalist bias leads its practitioners (with some notable exceptions) to regard these phonetic categories as ontologically suspect entities, whose exact nature is not worth the courtesy of clear thought. Very often, hope is expressed that all such subjective categories can be replaced by physical predicates of some sort, for instance by finding neurological signals of a sufficiently digital kind. The history of such efforts is not a hopeful one—I suggest that they should be abandoned, and that abstract representations should be granted the kind of status in phonetics that they are given in phonology or in cognitive psychology.

IV. Conclusion

I have suggested that the plan for an utterance is perhaps rather more abstract than traditional forms of phonetic representation, and that the process of speaking should be taken to account for at least some of the traditional data of phonological alternation. It follows that the process of realizing an utterance plan has at least some language-particular aspects, and cannot be entirely attributed to physics and universally determined physiology. Also, the fact that the process of speaking integrates prosodic and paralinguistic variation cannot safely be ignored; indeed, the study of what remains invariant under such variation can provide invaluable clues about the realization process and its linguistic inputs.

My suggestions may well be wrong; the true nature of the representation that underlies speech is obviously to be determined by research, as is the nature of the speaking process. The research in question is not well served by the traditional concerns and methods of either phonology or phonetics, and would proceed faster if the two scientific cultures were a little more like one another.

References

Liberman, M. *The intonation system of English*. Ph.D. dissertation, M.I.T. New York: Garland Press, 1979.

Liberman, M., & Pierrehumbert, J. A metric for the height of certain pitch peaks in English. *Journal of the Acoustical Society of America*, 1979, *66*, Suppl. 1, S64.

Liberman, M., & Pierrehumbert, J. Intonational invariance under changes in pitch range and length.(to appear).

Liberman, M., & Prince, A. On stress and linguistic rhythm. *Linguistic Inquiry*, 1977, *8*, 249–336.

Liberman, M., & Sag, I. Prosodic form and discourse function. *CLS*, 1974, *10*, 416–426.

Pierrehumbert, J. The perception of fundamental frequency declination. *Journal of the Acoustical Society of America*, 1979, *66*, 363–369.

Pierrehumbert, J. *The phonology and phonetics of English intonation*. Ph.D. dissertation, M.I.T., 1979 (forthcoming from MIT Press).

Pierrehumbert, J. Synthesizing intonation. *Journal of the Acoustical Society of America*,1981, *70*, 985–995.
Pierrehumbert, J., & Liberman, M. Modeling the fundamental frequency of the voice. *Contemporary Psychology*, 1982, *27*(9), 690–692.
Sag, I., & Liberman, M. The intonational disambiguation of indirect speech acts. CLS, 1975, *11*, 487–498.

Chapter 13

Some Reflections on Speech Research

FRANKLIN S. COOPER

I. Introduction

It was a privilege indeed to give the introductory paper at the Conference on the Production of Speech on which this volume is based. The topic is an important one, at the cutting edge of present-day speech research, so it is not surprising that several divergent paths are being followed. The meeting gave us an opportunity not only to compare recent findings but also to reexamine our research goals—to ask again what it is we are looking for.

In his letter of invitation, Peter MacNeilage suggested that I include a retelling—for his students, since the rest of the participants knew the story—of how Haskins Laboratories became involved in speech research and how the initial work on perception developed into parallel research on speech production. Since the story starts from a conceptual context that is no longer familiar or is but dimly remembered, it seemed useful to go back to the still earlier events and ideas from which acoustic phonetics emerged some 30-odd years ago. So, in the first half of this chapter, I have tried to cover very briefly the contributions of linguists and of engineers to concepts of speech that were current at the beginning of the fifties, and then to turn to events at Haskins Laboratories as a case history of how those concepts continued to evolve.

Who would not be tempted to push on from history to prognostication? I have tried to avoid the trap in the second half of the chapter and, instead, to look at present-day research from a little distance—to reflect on where it seems to be going and how this follows from current concepts about the nature of speech. In