Contents

Preface

This work is primarily addressed to phonologists interested in speech and to speech engineers interested in phonology, two groups of people with very different expectations about what constitutes a convincing, rigorous presentation. The subject matter, the application of autosegmental theory for Markov modeling, is technical, but not really esoteric – autosegmental theory is at the core of contemporary phonology and Markov models are the main tool of speech recognition. Therefore it is hoped that anyone interested in at least one of these two fields will be able to follow the presentation, and perhaps find something useful here.

As the title indicates, this is a rather formal work. There are formal theorems stated throughout the text, and readers who do not have a good background in calculus and linear algebra will have to take these on faith. On the other hand, readers with a science or engineering background will find the proofs (which are generally relegated to the Appendices at the end of each chapter) reasonably simple, even enjoyable. The main body of the text is basically self-contained. It should be easy to follow for everyone familiar with the basics of set theory, logic, and automata theory. All three topics are amply covered for example in Barbara Partee, Alice ter Meulen, and Robert Wall's *Mathematical methods in linguistics* (Kluwer Academic, Dordrecht 1990). Except for the Appendices, formalism has been kept to an absolute minimum, with arguments and even theorems presented in an informal, discursive style. Concepts are frequently introduced without a rigorous definition. In such cases their first significant occurrence is given in *italics* and when they receive a formal definition they appear in **boldface**.

Phonologists are advised to read the main text sequentially, and perhaps to ignore all the Appendices except for 2.5.3. In section 0.2 of the Introduction a chapter by chapter summary of the results is provided to aid the readers in devising a reading plan better suited to their interests. No knowledge of Markov modeling is assumed, but readers completely unfamiliar with the subject might want to consult L.R. Rabiner and B.H. Juang's "Introduction to Hidden Markov Models" in the January 1986 issue of *IEEE ASSP Magazine,* pp. 4-16, or the more extensive collection of papers in chapter 6 of Alex Waibel and Kai-Fu Lee (eds) *Readings in speech recognition* (Morgan Kaufmann, San Mateo CA 1990).

Speech engineers are advised to go from the Introduction directly to the last chapter, and work their way backward to the extent they wish to learn about the formal theory of autosegmental phonology that provides the motivation for the structured Markov models presented in chapter 5. There is an Index of Definitions, and many backward pointers are provided in the text to make this reading plan feasible. No knowledge of autosegmental phonology is assumed, but the reader interested in the linguistic motivation and use of the ideas which are studied in the thesis in a rather abstract fashion might want to consult John Goldsmith's *Autosegmental and metrical phonology* (Basil Blackwell, Oxford 1990).

Most of the material presented here is taken from the author's 1991 Stanford dissertation with only stylistic changes. The most important exceptions are sections 1.4.5, 2.5.4, and 5.3.6, which are intended to bring the reader up to date by providing critical assessment of subsequent work. Some parts of the material have been published or submitted for publication elsewhere: in particular, section 4.4 is now available in a self-contained version as "The generative power of feature geometry" in the *Annals of Mathematics and Artificial Intelligence* **8** (1993) 37-46.

Introduction

0.1 The problem

The last twenty years have witnessed a profound split between the engineering and the theoretical aspects of the study of human speech. In speech engineering, and in particular in speech recognition, these years brought the ascendancy of unstructured, statistical models over the structured, rule-based models. In the same period phonological theory came to emphasize the abstract, structural properties of sound systems over the directly observable properties of sounds, and created a highly algebraic theory that almost entirely ignores the variability of actual speech. This split is nowhere more clear than in the use of distinctive features: in speech recognition virtually no model uses features, while in phonology practically all research takes place in a feature-based framework. Is there a way to make such a massively justified and widely used theoretical device as features useful for speech engineers? Could phonology benefit from such an undertaking? This is the subject matter of this book.

Speech engineers and computational linguists crave after efficiency; they do not believe there has been an advance in the state of the art until they have seen a better implementation, a faster algorithm. Yet it is often the case that no amount of engineering ingenuity can push a given approach beyond some local optimum – what is needed is an entirely new approach, a conceptual breakthrough. The field of speech recognition is precisely in this state: for the last ten or fifteen years each advance in Markov modeling yielded increasingly diminishing returns, and the goal of creating systems that perform large vocabulary, speaker independent, continuous speech recognition with the same efficiency as humans is nowhere in sight. Where can a conceptual breakthrough come from? The present work grew out of the conviction of the author that for speech engineering the best source of new conceptual machinery is phonology. The approach taken here is to formalize autosegmental phonology in order to create a theoretically sound conceptual framework for speech recognition with Markov models.

Markov models offer an extremely powerful learning mechanism which is especially well suited for data with inherent random variability, but one that is in no way specific to the nature of speech data. Triphone models cannot exploit the large scale language-specific regularities of the speech signal, such as vowel harmony or root-and-pattern paradigms, and they do not scale up to pentaphones and even larger domains where these regularities would become accessible. Furthermore, standard Markov models create a structural split between phonetics/phonology (captured in the individual triphones) and morphology (captured in the lexical network connecting the triphones) while linguistic theory tells us that phonology and morphology are part of the same (stratal) organization and operate in an interleaved fashion that permits no split. Present-day phonology/morphology, though conceptually better equipped to deal with these issues, unfortunately does not provide us with a large body of well-defined and highly optimized algorithms that can be readily put to use in a speech recognition system – in fact it hardly provides any algorithms at all. In its present state, phonology is not ready for optimization, but it is ready for *formalization*: the key ideas, developed in the phonological literature in an informal fashion[1], can be expressed in a more rigorous manner so that the results can serve as the conceptual basis for algorithmization.

[1] Pullum 1989 characterizes the informal style used in contemporary phonology as follows: “Even the best friends of the nonlinear phonology that has driven the relatively formal pre-1977-style segmental phonology into the wilderness (...) will admit that it isn’t

0.2 The results

The most important overall result of this study is the creation of a model-theoretic framework that bridges the gap between the widely disparate practices of phonologists and speech engineers. Using this framework, the informally stated ideas of autosegmental phonology (AP) can be explicated, and the resulting model structures can serve as a blueprint in the design of speech recognition systems.

The syntactic devices used in expressing phonological generalizations are investigated in chapters 1 and 2, and the semantic interpretation of phonological representations is developed in chapters 3 and 4. The resulting model structures are then used as the basis of defining *structured Markov models* (sMMs) in chapter 5.

In the rest of this section the specific results are listed chapter by chapter and a brief discussion of their significance is provided. As can be seen from this list, the model-theoretic approach considerably improves the conceptual clarity of the often ill-understood technical devices used in phonological practice, and the design method stemming from it provides a completely new way of comparing and empirically testing a wide variety of specific proposals found in the phonological literature.

Main results of chapter 1

A. The notion "well-formed autosegmental representation" is rigorously defined (1.1-1.3, 1.5). Significance: forms the basis of all that follows.

---

trying to meet the conditions (...) for formal theories. True, a very significant outpouring of new ideas and new diagrammatic ways of attempting to express them has sprung up over the past decade; but it is quite clear that at the moment no one can say even in rough outline what a phonological representation comprises, using some exactly specified theoretical language. (...) Drifting this way and that in a sea of competing proposals for intuitively evaluated graphic representation does not constitute formal linguistic research, not even if interesting hunches about phonology are being tossed around in the process."

B. A linear encoding of autosegmental representations (AR-s) is developed. Significance: standard two-level software, originally developed for the linear case, can now be used for AR-s.

C. Asymptotic formulas are established for the number of well-formed, as well as for fully associated AR-s, and an exact relationship between the two series of numbers is established (1.6). Significance: solves known open problem of enumerating AR-s, gives exact measure of the information content of AR-s, provides the basis for D below.

D. The non-existence of optimal linear encodings is demonstrated (1.4). Significance: Results in B are shown to be near-optimal, hopes for totally eliminating autosegmentalization squashed.

Main results of chapter 2

A. The notion "well-formed autosegmental rule" is rigorously defined (2.1-2.2). Significance: completes the syntactic reconstruction of AP, paves the way for generative capacity result E below.

B. Phonological theories of rule ordering reconstructed in uniform framework of finite state control (2.1). Significance: Protects result E below against objections based on rule ordering.

C. Classes of autosegmental automata defined (2.3, 2.5). Significance: theory of automata and formal languages can be extended to ARs.

D. Encoding of multi-tiered representations investigated, basic method of synchronization presented (2.4). Significance: forms the basis of the reconstruction of synchronization in chapter 4.

E. Kleene theorem for bistrings established, finite-state-ness of AP demonstrated (2.5). Significance: extends classical result of Johnson (1970) to autosegmental phonology, forms basis of F,G below.

F. Variety of extant theories of reduplication explained in light of generative capacity (2.5). Significance: explains the reasons for the failure of the existing theories.

G. Obligatory Contour Principle explained as the limiting (simplest) case of a range of possibilities available in finite-state systems (2.5). Significance: puts debate on OCP in new light.

Main results of chapter 3

A. Klatt's deterministic model of duration reinterpreted as a probabilistic model predicting upshifted lognormal duration density (3.1). Significance: provides theoretical justification for C below.

B. Haskins Labs' deterministic model of duration reinterpreted as a probabilistic model predicting lognormal duration density (3.2). Significance: provides theoretical justification for C below and links the phasepoint/lag theory of synchronization presented in 4.2 to well-established phonetic theory.

C. Instead of the widely used normal model, a lognormal model of duration is proposed (3.3). Statistical proof of superiority of lognormal over normal obtained (3.3). Significance: lognormal provides a new, theoretically justified way of explicitly controlling duration density in semi-markov models.

D. The duration densities of the most important topologies of tied-state Markov models are found to converge to Dirac-delta (3.4.1-3.4.2). Significance: increased frame rate is shown to be disadvantageous for models without input probabilities.

E. Models with initial probabilities are shown to be trainable to fit any prescribed duration density distribution (3.4.3). Significance: replaces the complex probabilities used by Cox with real numbers in the [0,1] range, provides theoretical justification for input models.

F. Model structures containing random variables are introduced (3.5). Significance: the use of random variables is the key technical innovation needed for describing the meaning of ARs in a model-theoretic framework.

Main results of chapter 4

A. A general theory of features, based in natural classes, is developed (4.1). Significance: provides unified treatment of SPE, Pāṇini, and feature geometry, paves the way for E below.

B. The phasepoint/lag formalism of synchronization is introduced (4.2). Significance: provides the semantics for association lines.

C. Interval systems and interval structures defined (4.3). Significance: completes model-theoretic reconstruction of AP, forms the basis of sMMs presented in chapter 5.

D. Role of non-convexity and non-monotonicity in phonological theory investigated (4.3). Significance: underlying causes of non-monotonicity exposed.

E. Weakly boolean structures (Ehrenfeucht) are used to justify feature geometry (4.4). Significance: puts feature geometry in new light, makes relationship between contemporary and earlier theories clear.

Main results of chapter 5

A. Segmental interpretation is presented (5.1). Significance: provides the theoretical underpinnings for standard Markov models.

B. Cascade construction of sMMs introduced (5.2). Significance: captures the lack of synchrony among the features.

C. The possibility of training feature detectors is demonstrated (5.2). Significance: model need not rely on human expertise.

D. Recursive construction of sMMs according to a given feature geometry explained (5.3). Significance: enables linguist to choose between competing geometries on the basis of speech recognition performance.

E. Evaluation criteria for sMMs are presented (5.4). Significance: sMMs are a new class of Markov models, expected to be very successful in speech recognition. They are theoretically justified by AP, but unproven in practice.

0.3 The method

This work belongs in a broad scientific tradition, starting perhaps with Euclid, and probably best exemplified in modern linguistics by the early work of Chomsky, of using formal tools as a means of extending our knowledge about an empirical domain. In the first four chapters, the key ideas of autosegmental phonology are explicated[2], and in chapter 5 the resulting formal system is used for the construction of structured Markov models in order to link the actual practice of phonologists to the actual practice of

[2] The task of *explication* consists in transforming a given more or less inexact concept into an exact one or, rather, replacing the first by the second. We call the given concept (or the term used for it) the *explicandum*, and the exact concept proposed to take the place of the first (or the term proposed for it) the *explicatum*. The explicandum may belong to everyday language or to a previous stage in the development of scientific language. The explicatum must be given by explicit rules for its use, for example, by a definition which incorporates it into a well-constructed system of scientific either logicomathematical or empirical concepts. (...)

A problem of explication is characteristically different from ordinary scientific (logical or empirical) problems, where both the datum and the solution are, under favorable conditions, formulated in exact terms (for example. 'What is the product of 3 and 5?', 'What happens when an electric current goes through water?'). In a problem of explication the datum, viz., the explicandum, is not given in exact terms; if it were, no explication would be necessary. Since the datum is inexact, the problem itself is not stated in exact terms; and yet we are asked to give an exact solution. This is one of the puzzling peculiarities of explication. It follows that, if a solution for a problem of explication is proposed, we cannot decide in an exact way whether it is right or wrong. Strictly speaking, the question whether the solution is right or wrong makes no good sense because there is no clear-cut answer. The question should rather be whether the proposed solution is satisfactory, whether it is more satisfactory than another one, and the like. (Carnap 1950) speech engineers. No ink will be wasted on criticizing the lack of mathematical rigor in phonology, or the lack of theoretical orientation in speech engineering, as the author believes that more can be gained from trying to integrate the positive contributions of both fields than from trying to get people do things ‘properly’.

This emphasis on the positive contributions sets the present work apart from earlier attempts at developing a formal system of phonology and morphology. Categorial phonology (Wheeler 1981) and morphology (Hoeksema 1985), finite-state phonology and morphology (Kaplan and Kay ms, Koskenniemi 1983), or the more recent work on autosegmental phonology at Edinburgh (Bird and Klein 1990, Scobbie 1991) are certainly rigorous enough to satisfy even the most demanding taste. However, these systems do not offer a formal *reconstruction* of mainstream generative phonology, they offer formal *alternatives.* Because they explicitly reject one or more of the fundamental assumptions underlying the sequential mode of rule application used in the vast majority of generative phonological analyses, they do not make it possible to restate the linguists’ work in a formal setting – in order to enjoy the benefits of the formal rigor offered by these systems one must reanalyze the data.

The orientation of the present work is exactly the opposite: rather than championing the merits of any particular assumption, the aim is to create a meta-level formalism which is abstract enough to carry the often contradictory versions of AP as special cases. The definitions of well-formedness (section 1.3), rule ordering (section 2.1), rule types (section 2.2), HMM topologies (section 3.2), and feature geometries (section 4.1) are all made in this spirit. There are, to be sure, cases where the author cannot hide his sympathies completely, but the aim is to keep these to a minimum so that most autosegmental analyses can be faithfully replicated. It follows from this strategy that devices unique to a particular version of AP will not be analyzed in great detail; tools of the theory such as a reduplicative CVC template are not taken to be primitives but are built from the primitives supplied by the abstract framework. The advantage of this abstract outlook is that the work is not tied to any particular, and thus soon to be outdated, version of phonological theory.

Since the reader will not encounter sMMs until the last chapter, in a sense the bulk of this formal work is preparatory in nature. Given the rather wide-spread sentiment in speech engineering that linguistic models do not work and that it is altogether better to replace human intuitions about speech by automatically extracted knowledge (see e.g. Makhoul and Schwartz 1986), the question will no doubt be asked: why bother with all this theory? From the perspective of the speech engineer, the complexity of our preparations, and indeed the complexity of present-day phonological theory, can only be justified if it gives rise to more successful applications. But from the perspective of the phonologist the first four chapters are not preparatory at all; formalizing phonological theory is a worthwhile undertaking that can advance our conceptual understanding of language quite independently of its utility for speech recognition, speech synthesis, voice compression, speaker identification, or any other practical task confronting the speech engineer. The rest of this section discusses the logical structure of this undertaking, which is largely independent of the organization imposed by the specific results summarized in section 0.2 above. Readers more interested in the results than in broad metatheoretical considerations can skip this discussion without great loss.

What does phonological theory do? *How* does it do it? *Why* does it do it that particular way? These are the questions a detailed formalization should seek to answer. As for the first of these questions, most practicing phonologists view their theory as an instrument that will, much like the physician's X-ray machine, make accessible a well-defined part of the internal structure in humans that enables them to pursue a certain kind of activity, namely communication by means of conventional sounds or handsigns. And as an ordinary X-ray machine will bring into sharp relief the bones, and tell us little about the muscles, nerves, and other soft tissue equally important for the task of locomotion, phonological theory is focussed on a single component of communication, namely the *mental* *representations* associated with the sound/handsign aspect of the message communicated. Thus the first chapter is devoted to an explication of the mental representations assumed in contemporary phonological theory.

The second question, how phonology makes mental representations of the sound (or handsign) aspect of language accessible, is perhaps best understood from the perspective of writing and transcription systems. The move from mora-based or syllable-based to alphabetic writing systems introduces an abstract kind of unit that cannot be pronounced in isolation, namely (oral) stop consonants. The move from alphabetic to feature-based transcription (intimately linked with the early history of phonetics/phonology, see e.g. Jespersen's 1889 critique history of phonetics/phonology, see e.g. Jespersen's 1889 critique of Sweet 1880) results in completely abstract, unpronounceable units which embody the mental unity of articulatory and acoustic specifications (Jakobson, Fant and Halle 1952, Halle 1983). These units, and larger structures composed from them, can be made accessible via the study of the grammatical rules and constraints that are stated in their terms. Thus the second chapter is devoted to an explication of the rule and constraint systems used in contemporary phonological theory.

The third question, why phonology concentrates on the grammatical manifestation of mental units at the expense of their physical manifestations, has only a partial answer: the physical phenomena associated with speech are extremely complex, and their experimental investigation poses serious problems. As long as phonological derivations cannot be directly verified (because the nerve impulse patterns corresponding to the activation of mental units in the production and perception of spoken or signed language cannot be followed through the central nervous system), phonologists will have to rely on indirect evidence of some sort. But the difficulties in obtaining experimental evidence can only partially explain why contemporary phonology relies almost exclusively on grammatical evidence and why, in the rare cases when physical evidence is admitted, the articulatory domain is so strongly preferred.

The first major exposition of standard generative phonology, Chomsky and Halle 1968, devotes a full chapter to listing “the individual features that together represent the phonetic capabilities of man” but grounds the features only on articulatory correlates, mentioning “the acoustical and perceptual correlates of a feature only occasionally, not because we regard these aspects as either less interesting or less important, but rather because such discussions would make this section, which is itself a digression from the main theme of our book, much too long” (p 299). The most influential textbook of standard generative phonology, Kenstowicz and Kisseberth 1979, defines acoustic phonetics (p 7) but discusses only articulatory theory under the heading “linguistic phonetics” (pp 7-23). Expositions of the modern generative theory of features, such as Sagey 1986, again discuss articulatory, but not acoustic, evidence. Chapters 3 and 4 of this book are based on the view that the historical reasons for giving preference to grammatical over articulatory over acoustic data are no longer valid.

While it was certainly true a hundred years or even a few decades ago that careful observation of speech production yielded more reliable data than the “trained ear”, and that elicitation or introspection yielded even more reliable, quantized data about grammaticality judgments, neither of these points remains valid today. The recording and precise tracking of the position of the articulators during speech production is a major undertaking requiring specialized equipment of the sort described in Fujimura, Kiritani and Ishida 1973, while the recording and analysis of digitized speech can be performed on equipment no more complex than a personal computer. Furthermore, the inherently continuous and variable nature of speech data is brought under control by quantization and other modern statistical techniques, while the inherently quantized and invariable nature of grammaticality judgments becomes less and less pronounced as attention is shifted from the ideal speaker-hearer of the ideally homogeneous speech community to actual speakers in actual communities. Therefore, rather than excluding acoustic evidence from the domain of phonology, we should endeavor to create a "phonetic interpretation" that will map discrete phonological representations to physical events that unfold in real time.

The existing theories of phonetic interpretation, such as Keating 1988, Bird and Klein 1990, have two main shortcomings. First, they link phonological features to articulatory specifications and thus presume a thorough understanding of the relationship between the positions of the articulators and the acoustic signal. Second, they only describe the timing of (the beginning and end of) each gesture relative to (the beginning and end of) other gestures, but give no information about the absolute value of the time lags or the duration of the gestures. The theory developed in this book overcomes both of these shortcomings: it is applicable to all kinds of dynamically changing parameter vectors (be they articulatory, e.g. derived from X-ray microbeam records, or acoustic, e.g. derived by the kinds of digital signal processing techniques discussed in Rabiner and Schaefer 1979) and it is real time.

As a result of the work undertaken in the first four chapters, autosegmental phonology, and its phonetic interpretation, become a formal, readily algorithmizable theory of speech. However, it still suffers from a problem not much appreciated by linguists but taken very seriously by speech engineers: it is totally dependent on human expertise. In addition to the underlying representations and the rules, the grammarian will also have to specify the parameters of the interpretation. Since the number of such parameters is quite large, an automatic method of extracting them is clearly desirable. Chapter 5 is devoted to a new class of hidden Markov models which make it possible to perform parameter extraction (training) of phonologically motivated models using existing technology.

0.4 References

Bird, Steven and Ewan H. Klein 1990. Phonological events. *Journal of Linguistics* **26** 33–56.

Carnap, Rudolf 1950. *Logical foundations of probability*. University of Chicago Press.

Chomsky, Noam and Morris Halle 1968. *The Sound Pattern of English*. Harper & Row, New York.

Fujimura, Osamu, S. Kiritani and H. Ishida 1973. Computer controlled radiography for observation of movements of articulatory and other human organs. *Computer Biological Medicine* **3** 371–384.

Halle, Morris 1983. Distinctive features and their articulatory implementation. *Natural Language and Linguistic Theory* **1** 91–107.

Hoeksema, Jack 1985. *Categorial Morphology*. Garland Publishing, New York.

Jakobson, Roman, Gunnar Fant and Morris Halle 1952. *Preliminaries to speech analysis: the distinctive features and their correlates*. MIT Press, Cambridge MA.

Jespersen, Otto 1889. The articulations of speech sounds, Marburg.

Kaplan, Ronald and Martin Kay ms. *Phonological rules and finite state transducers*. Xerox PARC.

Keating, Patricia A. 1988. Underspecification in phonetics. *Phonology* **5** 275–292.

Kenstowicz, Michael and Charles Kisseberth 1979. *Generative Phonology*. Academic Press, New York.

Koskenniemi, Kimmo 1983. Two-level Morphology: a general computational model for word-form recognition and production, Department of General Linguistics, University of Helsinki Publication, Helsinki.

Makhoul, John and Richard Schwartz 1986. Ignorance modeling. In *Invariance and Variability of Speech Processes*, Joseph S. Perkell and Dennis H. Klatt, (eds.) Lawrence Erlebaum Associates, Hillsdale, NJ, 344–345.

Pullum, Geoffrey K. 1989. Formal linguistics meets the Boojum. *Natural Language and Linguistic Theory* **7** 137–143.

Rabiner, L. and R. Schaefer 1979. *Digital processing of speech signals*. Prentice Hall, Englewood Cliffs NJ.

Sagey, Elizabeth 1986. *The representation of features and relations in non-linear phonology*. PhD Thesis, MIT.

Scobbie, James M. 1991. *Attribute Value Phonology*. PhD Thesis, University of Edinburgh.

Sweet, Henry 1880. Sound notation. *Transactions of the Philological Society* **2**.

Wheeler, Deirdre W. 1981. *Aspects of a categorial theory of phonology*. PhD Thesis, UMASS Amherst.