Analytic models in phonology

András Kornai
IBM Almaden Research Center

---

0 Introduction

The goal of this paper is to describe a phonological framework in which the study of discrete 'phonological' units appears not as a separate enterprise, opposed to the study of continuous 'phonetic' phenomena, but rather as part of the larger enterprise of the scientific study of speech. Why is such a framework necessary? Given that the sound structure of language is a structure composed of discrete units defined in terms of contrast, would it not be the task of the phonetician to deal with all aspects of speech that are neither discrete nor contrastive? Unfortunately, phoneticians are not ready to pick up where the phonologists leave off and specify in detail how the discrete units used in phonology can be realized in, and recovered from, the undifferentiated continuous data provided by acoustic waveforms or articulatory records. In fact phoneticians have good reasons to be skeptical about the feasibility of the tasks imposed on them by the internal logic of phonology. For example, phonologists usually work with idealized data that preserves dialectally and grammatically conditioned variation but suppresses variation within the speech of a single individual and across individuals sharing the same dialect/sociolect. Perhaps the human apparatus for speech perception can somehow magically filter out variability along certain dimensions but not along others, but phoneticians have not succeeded in creating models that are capable of the kind of selective filtering presupposed in phonology, nor do they necessarily believe that humans actually perform such feats. The net result of the discrepancy between what phonologists assume phoneticians can do and what phoneticians actually do is that phonology depends on an unspecified and perhaps unspecifiable black box as its primary data-gathering instrument.

The result of this situation is that phonological theory is of surprisingly little use to anybody outside the self-imposed boundaries of the discipline. Phoneticians, acousticians, and other scientists and engineers interested in speech for theoretical or practical purposes generally find the rule or constraint systems devised by phonologists to be *brittle* in the sense that slight changes in the data, such as the discovery of a related set of forms, bring about radical revisions of the grammar, *fragmentary* in the sense that what constitutes relevant data for

the phonologist might not appear even in very large samples, while relatively frequent items often get ignored or simply deemed irrelevant, and hopelessly *handcrafted* in the sense that no general discovery procedure (learning from instances) exists.

Our primary design goal in this paper is *explicitness:* we wish to create a system that in no way depends on the output of unspecified modules. Our secondary design goals are *robustness*, the ability to adapt the system to slightly different data sets without major changes in the grammar, *completeness*, the ability to deal with actual data sets without externally emphasizing or de-emphasizing parts of them, and *automatic acquisition* that sets up the structure (including the units of representation) without relying on human expertise. None of these goals are incompatible with current phonological theory, and this paper does not advocate any radical departure from the central ideas of generative phonology beyond the argument that as phonologists we must take full responsibility for the empirical grounding of our theories. Since we can not simply hand over certain tasks to phoneticians and pretend that they are willing and able to deal with them, we must do the job ourselves. It should be emphasized at the outset that the issue is not a definitional one, whether the resulting theory should still be called phonology in spite of its partially phonetic character, or whether the researchers doing this work should be called phonologists or phoneticians – the issue is how to do the work.

Section 1 sets the stage by presenting a broad view of underlying and surface representations and distinguishing *combinatorial* and *analytic* theories in phonology. Section 2 provides an introduction, specifically aimed at the phonologist, to the basic statistical methods of dealing with variability in the data. Because the fundamental techniques for coming to grips with raw speech data were developed in the speech engineering community, well outside the mainstream of phonetics, it is hoped that most phonologists, even those familiar with the phonetics literature, will find the discussion of these methods to be of some interest. Finally, Section 3 presents the formal model developed in Kornai (1994) in a relatively informal fashion, and describes how in this model the traditional *intersective* theory of features can be replaced by a *superposition* theory. Throughout the paper, the emphasis is on the external (methodological) justification, the internal logic, and the overall goals of the theory – formalism is kept to the absolute minimum and no mathematical sophistication is assumed on the part of the reader

1 Underlying and surface representations

Any theory of phonology makes an ontological commitment to some, possibly infinite, set of psychologically real units, though not necessarily to the way in which complex units are constructed from simpler (atomic) units. Even if com-

plex units are specified in terms of their constituents as a matter of convenience, there need not be a commitment to the psychological status of this specification procedure. Intermediate representations, procedures, and structures can be viewed as having no theoretical status whatsoever, comparable to the scratch paper that holds the intermediate results in long division, or they can be treated as significant, depending on the theory. For example, Chomsky and Halle (1968) are not committed to the reality of the intermediate stages of an SPE-style derivation, Koskenniemi (1983) to the individual rules that combine in a single finite state transducer in two-level phonology, nor Prince and Smolensky (1993) to the suboptimal forms discarded in optimality theory, but level-ordered theories of phonology (Kiparsky 1982, Mohanan 1982, Booij and Rubach 1987) generally treat the output of the the individual levels as real.

In addition, any theory of phonology must rely on some source of empirical data which is external to the theory itself. While interpreting the raw data might require considerable phonological sophistication, there is a methodological difference between the interpretation process, typically viewed as being part of the theory, and the data gathering process, which is typically viewed as being outside the scope of the theory. For our purposes, empirical data is best classified in terms of how much the human apparatus for speech production and perception is used as an instrument in collecting the data. At one extreme, we find those theories that rely exclusively on subjective data, typically judgments concerning the grammaticality and well-formedness of certain forms. At the other extreme we find theories using only objective data such as speech waveforms or direct measurement (nowadays generally by the X-ray microbeam method of Fujimura et al. 1973) of the movement of the articulators.

1.1 Combinatorial theories

To a surprising degree the ontological and the empirical basis of a phonological theory can be inferred from, or even identified with, its notions of underlying and surface representations, independent of whether these representations are linked to one another by declarative or procedural means. Most theories of phonology, including atomic (van der Hulst 1989), autosegmental (Goldsmith 1976), constraints and repairs (Paradis 1988), declarative (Scobbie 1993), dependency (Anderson and Ewen 1987), government (Kaye et al 1985), and lexical phonology (Kiparsky 1982), are concerned primarily with the relationship between mental representations built from discrete, atomic primitives and a broad transcription of the speech signal. Since both the primitive theoretical units and the empirical data are assumed to come from the same, discretely generated set, specifying the relationship between the two is a matter of (infinite) combinatorics, and we shall call all such theories *combinatorial.* The principal mathematical tool used in combinatorial theories are the systems of context sensitive rewrite rules introduced in Chomsky (1956) and elaborated in SPE, but other algebraic specifica-

tion methods such as categorial grammars (Wheeler 1981), logic programming (Bird 1990), rational relations (Kaplan and Kay 1994), and constraint satisfaction (Scobbie 1991, Bird and Ellison 1994) have also been used. Optimality theory, relying on a highly restricted version of default logic, also belongs here.

Combinatorial theories presume a rather sophisticated view of the mental lexicon, assuming that it stores highly preprocessed discrete units, and an equally sophisticated view of the realization process, assuming several stages of central processing (application of phonological rules/constraints) resulting in complex nerve impulse patterns driving the articulators, with the final output determined by the acoustics of the vocal tract. It is precisely the complexity of these assumptions, and the dearth of predictive theories about the individual stages, that leaves phonology in the uncomfortable position of having to depend on black boxes for its input data. As a case in point, let us consider the theory of distinctive features (Jakobson 1939). A rather detailed qualitative description of the articulatory and acoustic correlates of distinctive features was available as early as 1952 (Jakobson et al 1952). Nearly three decades later, Stevens and Blumstein (1981) still had not found a way of turning this into a quantitative description that could be used to automatically detect features. Though Halle (1983) reiterated the commitment to a direct neuro-biological interpretation of features, the black box remained impregnable. To this day, in spite of repeated efforts such as Cole et al (1983, 1986), research in this area has failed to reveal a set of reliable acoustic cues for phonological features of the sort envisioned in Cherry, Halle and Jakobson (1953) and Cherry (1956).

Sixty years of failure on the part of phonetics to supply phonology with a suitable interface can not easily be explained by some kind of institutional hostility to Prague Circle ideas. While such an explanation might have contained a grain of truth in the fifties and the sixties, by the seventies a great number of phoneticians were sympathetic to generative ideas, and in the last decade tremendous progress has been made in relating modern theories of phonological representations to articulator movement (Keating 1985, 1988, Browman and Goldstein 1987). But in one crucial respect these models inherit the weakness of the original Jakobson-Fant-Halle model: they are still qualitative, and offer no upward path to a quantitative theory. There is a long way from observing the waveform to inferring the position of the articulators (a feat central to the motor theory of speech perception, Liberman et al 1967), but even if this could be done, or even if an X-ray movie is provided, we are still not capable of automatically detecting which sound or feature is being produced. Therefore in the work reported on here we present a more integrated view of speech structure, one that does not depend on unspecified theory-external modules for its empirical data.

1.2 Analytic theories

The simplest and most direct view of the mental lexicon is to assume that it stores highly specific acoustic engrams recorded during the language acquisition process, and these engrams can be directly used as lookup keys into a mental database that will contain syntactic, semantic, morphological, and other non-acoustic information about the form in question (Klatt 1970). Under this view, surface forms would be acoustic waveforms, while underlying forms could contain detailed articulatory plans for the production of the form, together with links to semantic, syntactic, and morphological information stored in various formats. We will take this model as our starting point, and argue that any departure from a simple and direct model must be justified by a great deal of evidence. We will see cases where such evidence can indeed be cited, but in general our methodological stance is a conservative one in which entities are not multiplied unless necessary.

Let us first consider the issue of discrete units. Since both acoustic engrams and articulatory plans can be stored as distributed representations (e.g. by adjusting connection strengths in a neural net), there is no particular reason to suppose that the underlying forms are discrete. If anything, experience with artificial neural nets suggests the null hypothesis that everything is continuous unless proven discrete. Neither memory nor the machinery of speech production lend themselves to a thorough characterization in terms of a small number of discrete states. To be sure, certain physical properties, such as whether airflow is laminar or turbulent, do lend themselves to such a characterization. But for the most part, we have a continuous system: within the boundaries dictated by anatomy, articulators are free to assume any configuration in a multidimensional continuum. The same is true of the acoustic signal: it is striking how few category distinctions can be established by direct inspection of the signal, and how fluid the boundaries between the categories are. Finally, hard-wired perceptual categories can seldom be found. Even to the extent we find categorial perception (Repp 1983), we find it mostly based on acquired, rather than inherited, distinctions.

In certain cases biology, physics, or cognitive science may furnish some mechanisms external to phonology to implement, or at least to motivate, some form of discretization. But there is a whole range of cases such as tonal levels, voice onset time, modes of phonation, or vowel articulation, where quantal effects cannot be attributed to external mechanisms of any sort – rather, it is only the discrete nature of the linguistic signal that imposes some rudimentary discretization on the continuous physical signal. Yet we find a continuous encoder, channel, decoder, and memory harnessed to the goal of generating, carrying, understanding, and storing messages composed of discrete units. How this can be done is the central question of a wide range of models from Stevens' (1972) quantal theory to contemporary theories of speech recognition (starting with

Baker 1975), theories that we will call *analytic*, since their primary formal tools are mathematical analysis and statistics. Laboratory phonology (Kingston and Beckman 1990), the investigation of phonetic phenomena from a direct phonological perspective, also belongs here.

2 Statistical models of underlying units

In order to characterize the relationship between psychological units of linguistic processing and their physical realizations we need a tripartite characterization. First, an inventory of psychological units must be presented, and the valid combinations of the psychological units must be enumerated and represented in a symbolic system. Second, we need a method for measuring and reproducing utterances in a mechanical fashion. Finally, we need to specify which utterance corresponds to which representation under what conditions. The first task is addressed, as we have seen, by all combinatorial theories of phonology. The second can be performed by a number of instruments: for the sake of concreteness we assume that the waveform has been recorded and digitized at a sample rate and quantization comparable to that of standard music CDs. As for the third, we obviously have a many to many relation, with different phonological representations corresponding to the same utterance (neutralization) and the same representation having many realizations, and many conditioning factors, ranging from the inevitable physical differences among speakers sharing the same competence to the amount of distortion tolerated in the realization process. In 2.1 we will summarize the principal ways of arguing for a given inventory element on the basis of grammatical or extragrammatical data. In 2.2 we discuss how traditional linguistic criteria used in establishing some element can be reformulated as statistical criteria.

2.1 Justifying the primitive units

Analytical theories do not necessarily share the ontological commitment of combinatorial theories to discrete primitives such as features, phonemes, morphemes, or words. Some of these units, most notably the phoneme, are instrumental in describing such a broad range of phenomena that their psychological reality can hardly be disputed. But other widely used units, e.g. the abstract quanta of stress (asterisks) employed in a variety of grid-based theories of stress (Halle and Vergnaud 1987), are much more restricted in their usefulness and indeed their very existence as units of mental representations can be questioned. Without some radical improvements in neurophysiological instruments neither memory engrams nor nerve impulse patterns can be directly investigated at the level of detail required for the study of linguistic processing. When (and if) such revolutionary technology becomes available it will be possible to investigate these

units by directly tracing the causal chain responsible for their emergence in continuous media. Until then, we are forced to accept less direct methods of proof, to which we turn now.

The first argument in favor of the existence of a particular unit is the introspective one: most researchers are convinced that they are in fact communicating using sentences, words, syllables, and phonemes. A great deal of the reluctance of speech engineers to accept distinctive features can no doubt be attributed to the fact that for features this argument fails: no amount of introspection reveals, say, the featural composition of vowels.

The second argument in favor of certain linguistic units can be made on the basis of particular systems of writing. To the extent that a morpheme-, mora-, syllable-, or phoneme-based writing system can be easily acquired and consistently used by any speaker of the language, the psychological reality of the units forming the basis of the system becomes hard to deny. Distinctive features fare slightly better under this argument, given the Korean alphabet and some early sound-writing systems such as Bell's (1867) Visible Speech, but to make the point more forcefully the ease of use and portability of such writing systems to other languages needs to be demonstrated.

Finally, the favored mode of argumentation in linguistics is based on the economy introducing a particular unit brings to the description. It is not entirely obvious that without independent support such arguments are strong enough to establish the psychological reality of a unit. While there is no doubt that this kind of structural evidence massively favors the use of distinctive features, more direct psycholinguistic experiments bearing on the psychological reality of features are far from conclusive (Remez 1979).

2.2 The statistical justification of units

For analytic theories of phonology, the problem of justifying units reduces to the problem of recognizing instances in the raw (articulatory, acoustic, or auditory) data. This makes it possible to bring another kind of evidence to bear: in addition to the purely extralinguistic and the purely grammatical forms of evidence discussed in 2.1 above, we can also consider direct statistical evidence. The traditionally central criteria for establishing phonological units, *recurrence, distinctiveness,* and *parsability* can be reformulated as conditions on a *canonical target* model, which, though never spelled out in formal detail, is implicit in most applied work dealing with speech synthesis and recognition. Let us consider each of these criteria in turn.

RECURRENCE AMOUNTS TO THE ASSUMPTION OF A DENSITY PEAK IN A CONTINUOUS DISTRIBUTION. The classic Peterson-Barney (1952) data on vowel formats will, even with the phoneme labels removed, present a picture very different from a random set of dots in 2- or 3-dimensional space. (The first three formants are available in the Peterson-Barney data, but in the figures

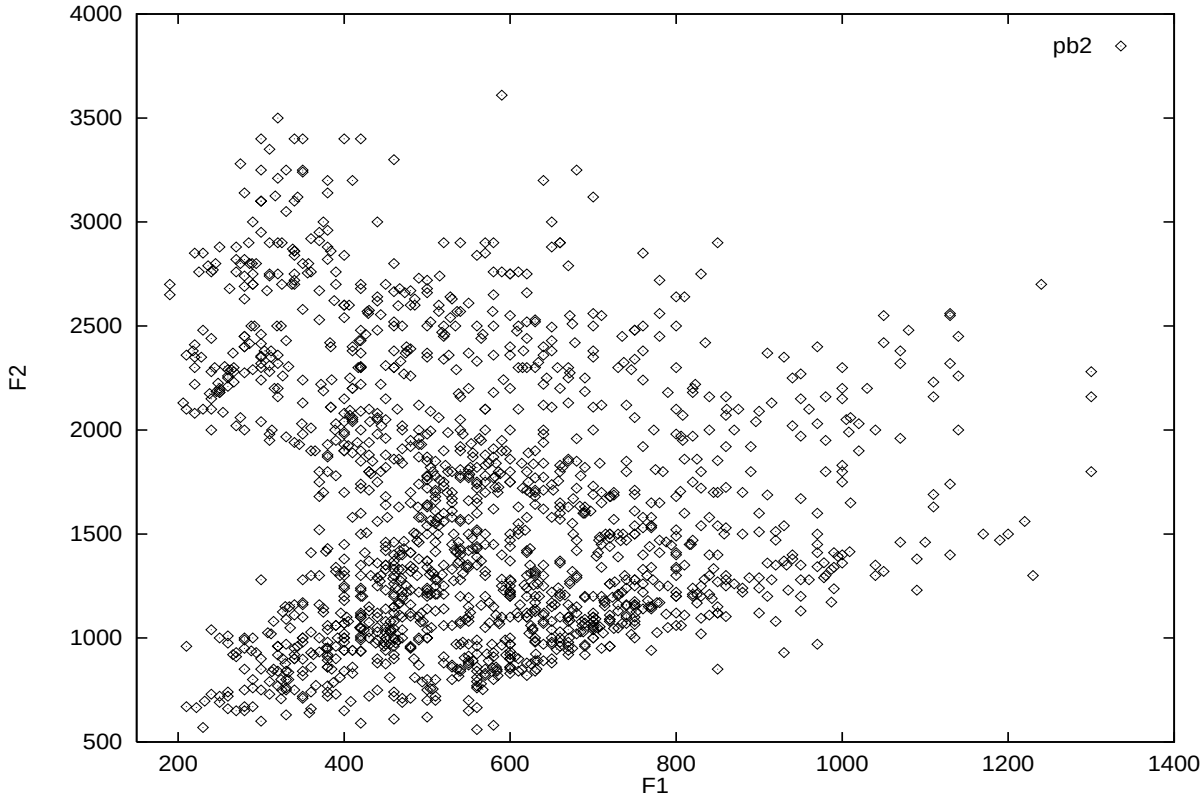

Figure 1: Vowel formants (after Peterson and Barney 1952)

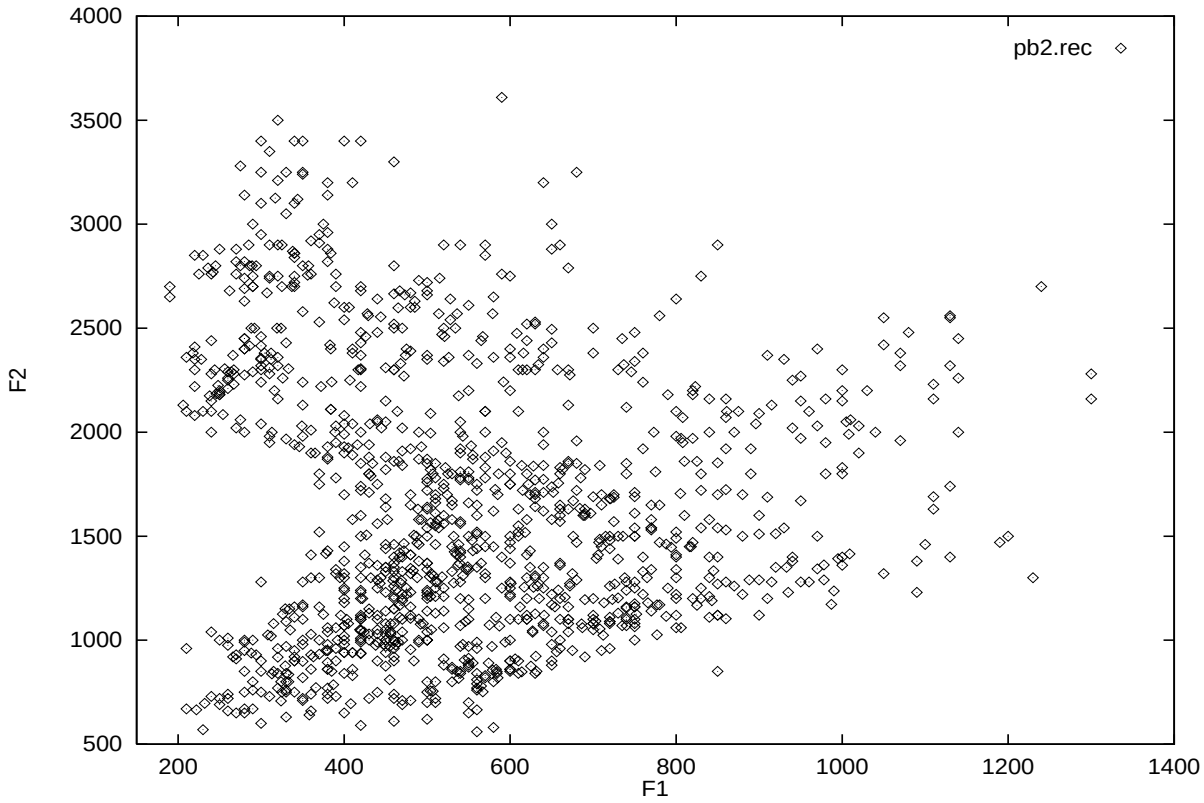

Figure 2: Unambiguous tokens produced by adult male speakers

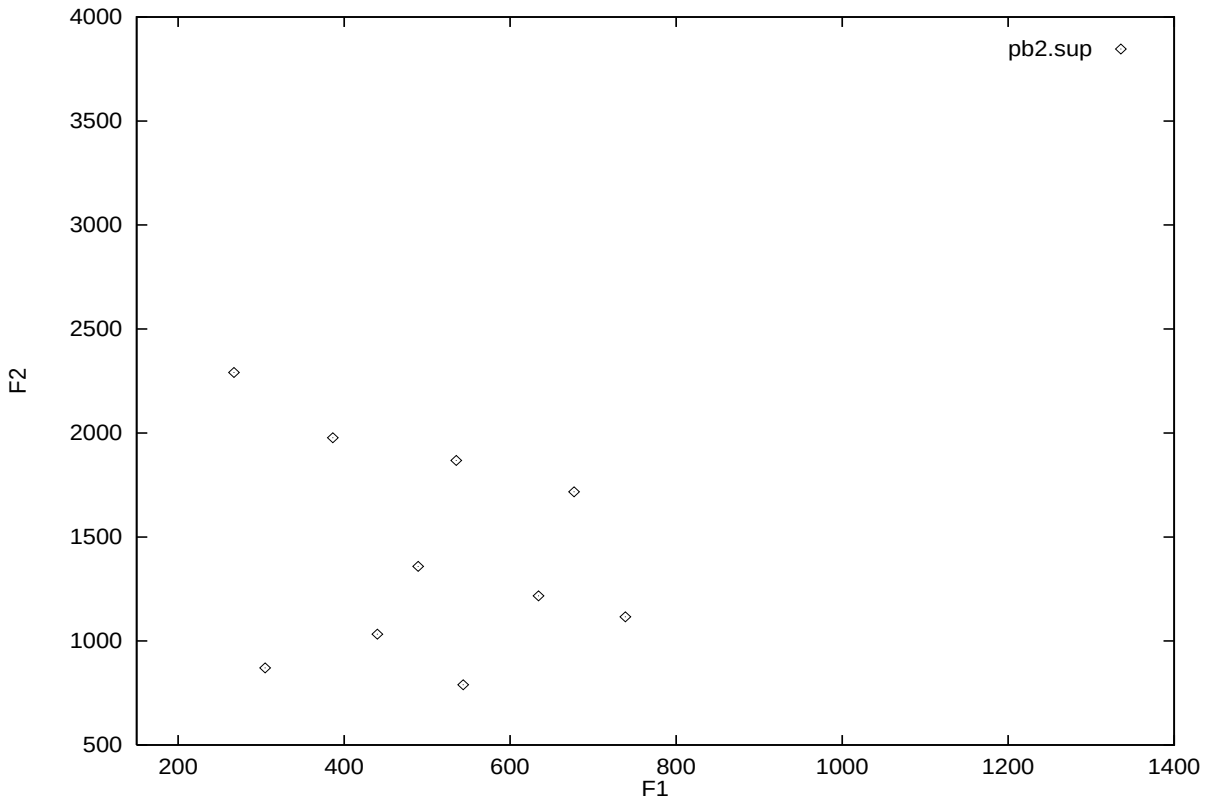

Figure 3: Mean values computed for 10 phonemes

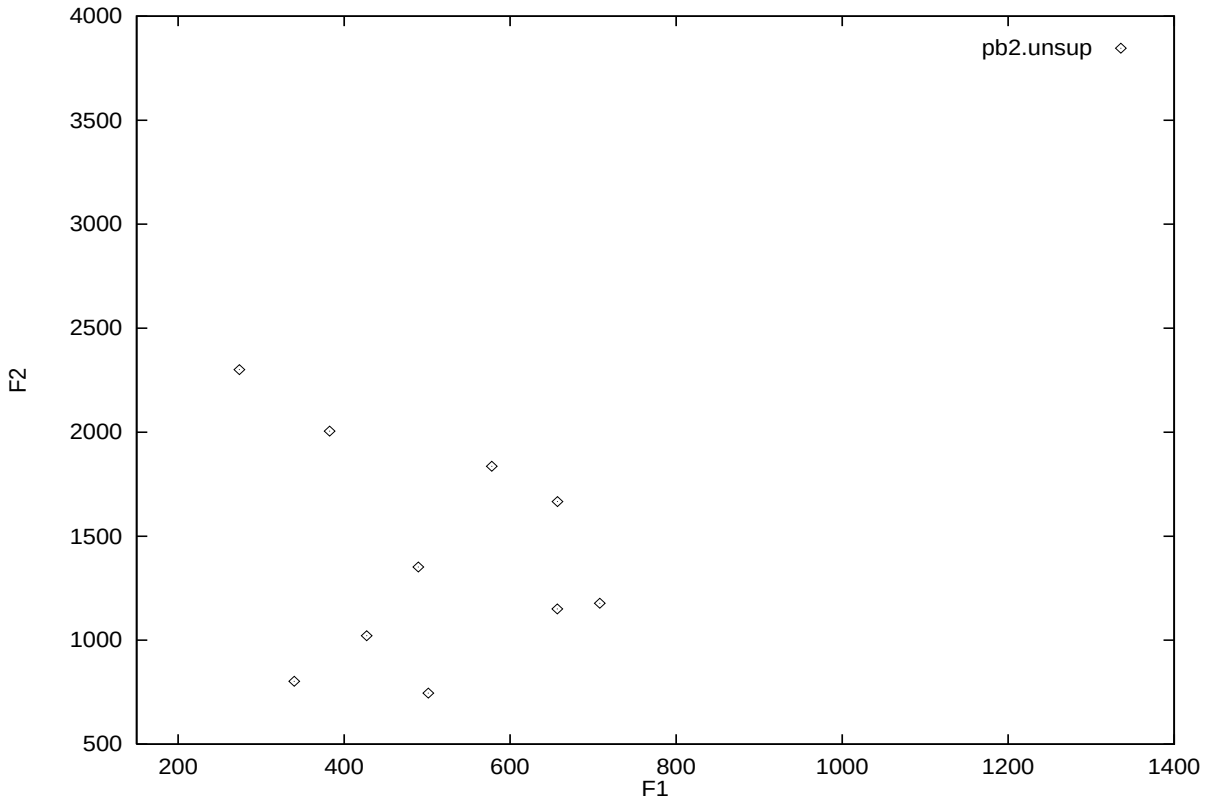

Figure 4: Mean values computed for 10 best clusters

only the first two are displayed). Because the dots corresponding to the vowels appear in a limited region of the plane (Fig. 1), their distribution can not be uniform. Even within the region of interest, which is obviously constrained to be rather small by the acoustic parameters of the vocal tract, the dots are spaced more densely in some regions than in others: simple visual inspection and more sophisticated statistical criteria both show the data to be clustered around relatively few points. If we remove all data points that correspond to utterances listeners could not identify, and include only the adult males in the sample, this clustering tendency is even more visible (Fig. 2). When the phonemic labels are available we can use *supervised* clustering techniques (Anderberg 1973) to capture this fact, and to compute statistics, such as the mean formant positions, that characterize the ten phonemes considered by Peterson and Barney (see also Watrous 1991) in direct acoustic terms (Fig. 3). But if phonemic labels are not available, *unsupervised* clustering techniques (Everitt 1980) are still applicable: we can simply compute the ten best clusters and see what formant positions they will correspond to. Figure 4 shows the results of this computation with the standard k-means algorithm, run in three dimensions, starting with fifty random clusters and ending in ten.

From the perspective of the phonologist, the most surprising idea here is not so much that generic statistical techniques can successfully reduce the complexity of the data, but rather the assumption of a direct mapping between psychological units and waveforms. In analytic models phonemes are defined extensionally, as types defined by the collection of their tokens, and not by some inherent (intensional) properties such as voicing or tongue position. The distinction between underlying and surface units is replaced by a distinction between populations and samples. In ideal cases, as in repetition tasks where most sources of variability are controlled, we find very tight clusters for each phoneme or major variant (e.g. tapped or trilled r) and we can basically characterize each population by a single sample. When variation is small, we can think of the average of the cluster as providing a canonical value, distorted only by imperfections in the muscular micro-control of the speaker. In less ideal cases, the variation will be larger, and characterizing any population will require more numerical information than what is provided by the mean. But intensional characterization, e.g. in terms of distinctive features, can be justified only to the extent it leads to better quantitative characterization. (A distinction should be made between the statistical norm, defined by the mean, and the community norm, defined by some acoustic/auditory ideal shared by the speakers. This distinction is likely to play an important role in long-term processes such as vowel shifts, but for the purposes of describing synchronic phenomena can be ignored for the most part.)

In the mathematically simplest version of canonical target models, the populations are modeled as normal (Gaussian) distributions. To characterize an n-dimensional normal distribution we need two sets of parameters: the means, which give the coordinates of the Gaussian peak, and the covariances, which de-

scribe the size and shape of the ellipsoidal region around the peak that contains the bulk of the data points. As a further simplification, the axes of this ellipsoid are often assumed to be parallel to the coordinate axes: in such *diagonal covariance* models each canonical target can be given by $2n$ parameters. *Full covariance* models require $n(n+3)/2$ parameters, still a very small number if we consider that each data point is characterized by n parameters (e.g. the first n formants) and each cluster is actually built on thousands of data points (and potentially subsumes an infinite number of them). While the mathematically simplest model is not necessarily the best one, normal distributions are widely used in practice, and make it easy to explain the statistical meaning of the distinctiveness criterion for establishing phonological units, to which we turn now.

DISTINCTIVENESS AMOUNTS TO THE SEPARATION OF DENSITY PEAKS. Again we take the unsupervised case as our starting point, and again we appeal to properties of the data already accessible through direct visual inspection but more solidly established by statistical tests: whether the data is best described by a single density peak or several distinct peaks (Fig. 5). Wherever we find a distinct peak, a distinct unit is justified at some systematic level. This need not be the 'systematic phonetic' level: for example, the first three peaks in the frequency distribution in Figure 5 are a systematic effect of having men, women, and children in the sample. Sometimes, as in the case of tapped vs. trilled r, distinct peaks can be shown to be subsumed under a single abstract unit, but more typically statistical distinctness implies psychological distinctness. The burden of proof is always on the proponent of the more abstract unit. We should emphasize here that distinctness of clusters is a type-level property, not to be confused with distinctness of tokens. In labeled data (Fig. 6), where the phonological value of each data point is known, we will often find data points with different phonological values falling quite closely to one another (neutralization) and we will also find distant data points within the same phoneme (allophonic variation). Both of these cases are quite compatible with the notion that the density peaks characterizing the different clusters are placed at considerable distance from one another: neutralization means that the tails of different distributions can overlap, and large allophonic variation means that the (co)variances are large i.e. the distribution is not very peaked.

To accommodate the cases of major allophonic variation, often several distinct Gaussians are assigned to a single phoneme model: in this case we talk about *mixture models* because the density function describing the distribution of data points belonging to a single phoneme is a mathematical mixture (weighted sum) of ordinary Gaussians. This method opens the door to adding an inordinate number of mixture components as an expedient way of achieving better fit with the data: in the limiting case, we can fit a very narrow Gaussian to each data point and thereby achieve perfect fit. However, the number of Gaussians that can be justified is limited by the Minimum Description Length principle (Rissanen 1978), and parametric models of speaker-dependent characteristics will in

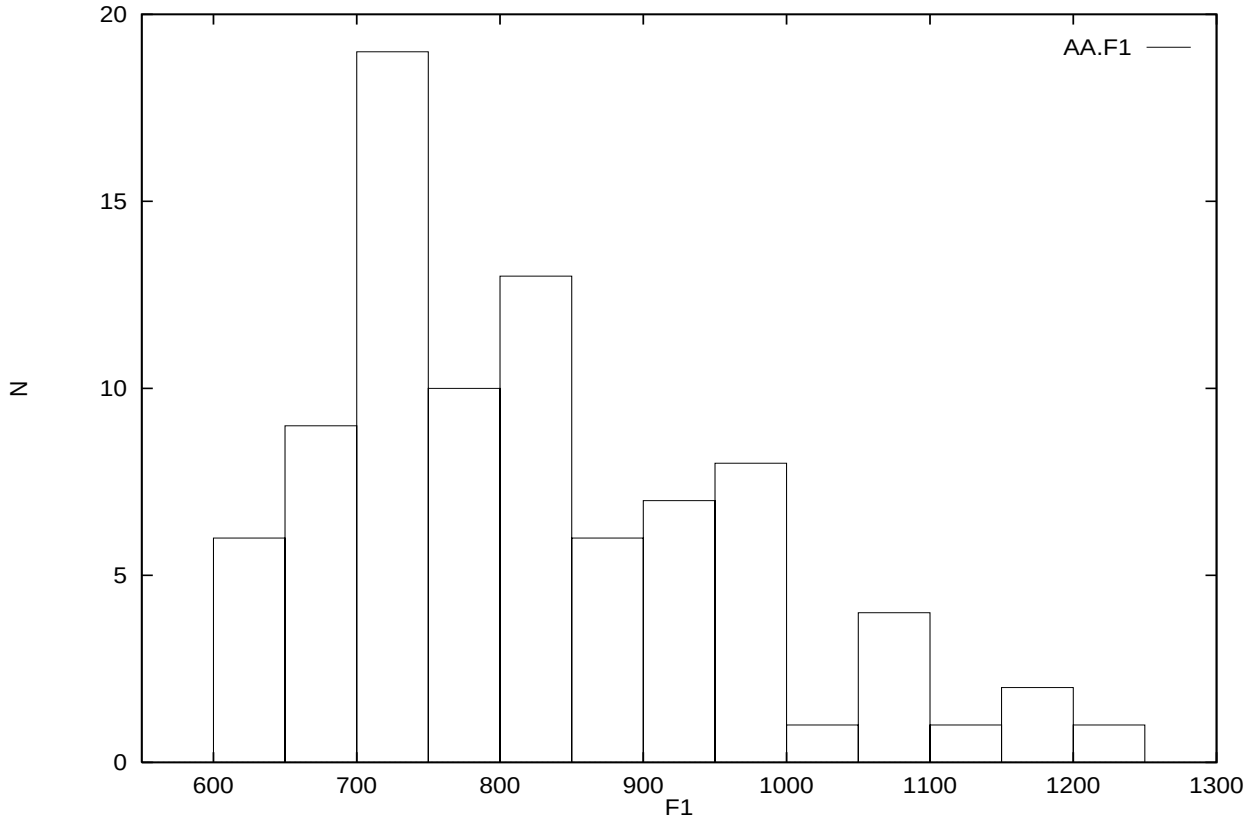

Figure 5: Distribution of F1 data for AA

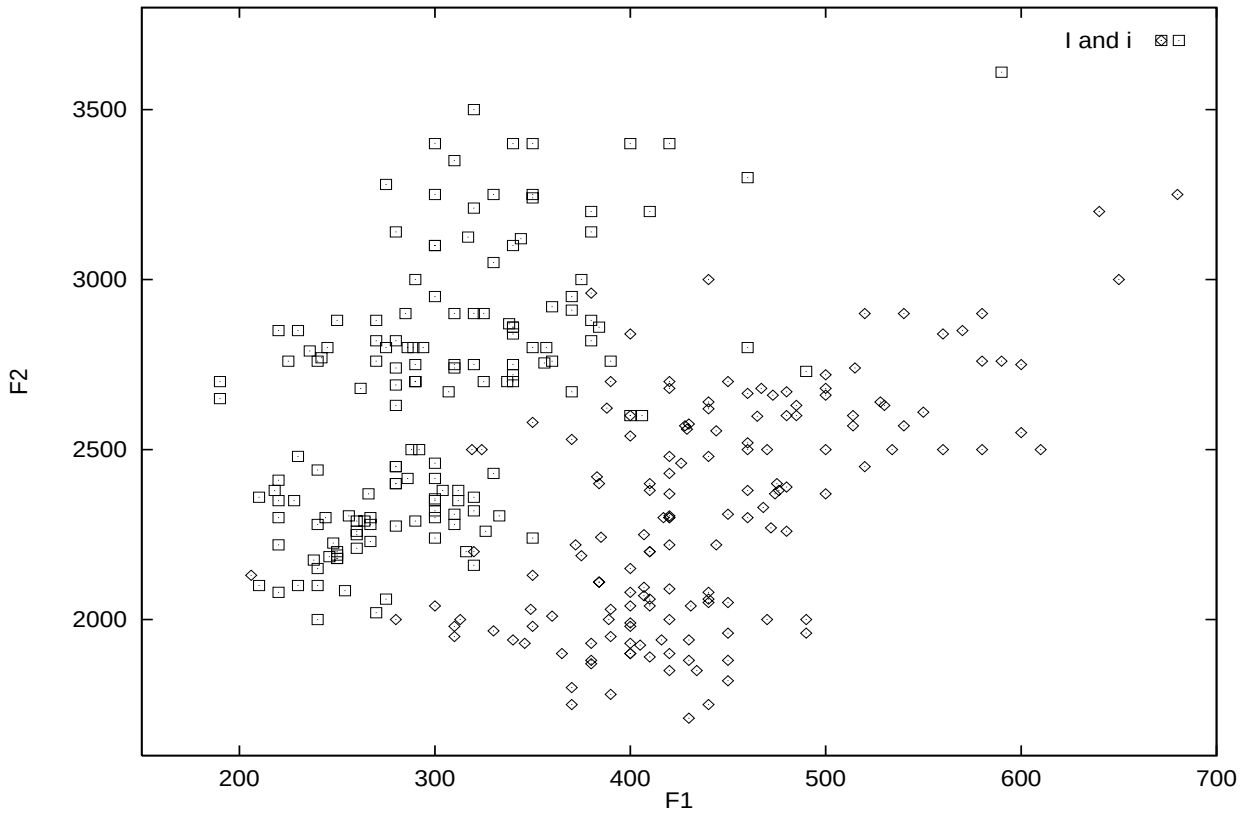

Figure 6: Unambiguous tokens produced by adult male speakers

general reduce the size of mixtures much better than mechanical curve-fitting.

PARSABILITY AMOUNTS TO A CONDITION OF APPROXIMATION BY INTERPOLATION. This criterion comes into play when we shift our attention from quasistationary signals and the paradigmatic relations that obtain among them to dynamic signals and their syntagmatic relations. Monophthongal vowels and liquids provide ideal examples of quasistationary signals, and their characterization in terms of time-invariant measurements such as formant place and intensity is quite reasonable, both for synthesis and analysis. For an utterance composed entirely of such sounds, the complex trajectories described by the articulators and/or by the features used to characterize the signal, it is sufficient to describe the quasistationary values that obtain during the central portion of each sound: the rest can be interpolated (see 3.2).

Unfortunately, no time-invariant characterization has ever been provided for the majority of the cases, particularly for stop consonants, nor is such a characterization likely to be forthcoming. It is not impossible to characterize dynamic systems in static terms, for this is exactly what reference to (higher) derivatives accomplishes. But the system under investigation is so highly dissipative that bringing the Hamiltonian apparatus of classical physics to bear is a nontrivial problem. On the other hand, stop consonants can be described with relative ease as a succession of states (e.g. the lips closing and then bursting open), so the difference between static and dynamic gestures is merely a difference in the number of freeze-frames we need for a full characterization: monophthongs require one, diphthongs require two, stop consonants might require three or even more. Keeping this in mind, the idea of canonical target configurations and interpolation is still applicable, but with the important caveat that a single phoneme might encompass several canonical states or *microsegments* (Fant 1973).

Once we divorce the idea of freeze-frames from the idea of minimal concatenative units, as suggested by monophonemic analyses of diphtongs and affricates, the phoneme is no longer the only candidate for underlying unit. Syllables, moras (demisyllables), and other units for which concatenation-based speech synthesis models exist (for an overview, see Klatt 1987) can also be realized as a succession of freeze-frames. Even models based on units of highly questionable psychological reality, such as diphones (units composed of the second half of one phone and the first half of the following phone) work reasonably well. The logic of our enterprise dictates that we treat all candidate units, including the implausible ones, equally, and select the best candidates according to statistical criteria based on the ability of the proposed unit(s) to reduce the complexity of the data.

It should be emphasized again that adopting an analytic model is not the same as denying the psychological reality of phonological units. To the contrary, the analytic model rests on the assumption that psychologically real underlying units exist, and can be extensionally defined by the statistical ensemble of the corresponding speech intervals. Whether such ensembles should be di-

rectly characterized by their means, variances, and other statistics, or whether they are better characterized indirectly, by causal models tracing the chain of production and/or perception, is ultimately an empirical matter.

3 Similarity

The canonical target model, implicit in most analytic theories of phonology, has been explicitely formalized in Kornai (1994), a paper aimed at computer scientists and mathematicians who can be expected to use formalism as their guide through largely unfamiliar empirical terrain. Here, aiming at a reader familiar with modern phonology but not with advanced mathematics, we describe the genesis of canonical targets starting from general methodological considerations and often informal linguistic models. In 3.1 we describe the model using phonemes or similar concatenative units, and 3.2 describes a version using features or similar nonconcatenative units.

3.1 Targets and interpolation

It is rather unfortunate that phonologists interested in the study of speech are often stopped from making their own contributions by the sheer complexity and mathematical sophistication of the signal processing techniques routinely used by speech engineers. Here we abstract away from the exact nature of the signal processing and treat all kinds of formant tracking, filter-band analysis, linear predictive coding, cepstral analysis etc. (for an overview see O'Shaughnessy 1987) as some transformation T that can be applied to a waveform to reduce its size. The idea behind transformations is that they are fully invertible in the limiting case: if no higher coefficients are truncated, we get lossless compression. If the higher coefficients are truncated to save space (or not even computed to begin with) the compression will be lossy. However, to the extent that synthesis based on transformed data provides results that are not perceptibly different from the original waveform, we can still think of these methods as invertible.

From our perspective, the main point is not so much the reduction in data size as the ability of transformations to turn near-periodic waveforms into near-constant functions. The price we pay for this ability is that instead of a real-valued (scalar) function of time we must use n-dimensional (vector) functions, which correspond to *trajectories* of a point in n-dimensional real or complex space. Recovery of articulatory parameters, to the extent feasible, is just another transformation that replaces the fast changing waveform by a set of more slowly changing parameters. This particular black box can be penetrated by X-rays, and readers more comfortable with physical transformations than with mathematical ones can think of the trajectories that we will discuss as physical trajectories of the articulators.

Unlike the transformations generally used by engineers which are easy to compute but hard to interpret, articulatory parameters are hard to compute but have an obvious physical interpretation. We would consider the computational effort well spent if we could bring the physical laws governing articulator movement into play, but, unfortunately, the dynamics of articulator movement is not very well understood. Though with the advance of finite element techniques we can expect significant progress in this area, it should be kept in mind that the most recalcitrant aspect of speech data, variability, is in no way reduced by moving from the acoustic to the articulatory domain. Whatever the eventual form of the physical model will be, it will contain a large number of stochastic parameters for the size of the vocal tract, muscular strength of the individual, and other biological factors. While we can hope for important insights from this area, especially in the separation of intra- and inter-speaker variation, progress in phonology does not depend on progress in physical modeling of the vocal tract. To the contrary, articulatory models of analysis and synthesis have a long way to go before their performance becomes comparable to that of models using more abstract mathematical transformations such as cepstra. The relevance of the study of the articulators for phonology in general, and sign language phonology in particular, remains to be demonstrated.

In sum, we can treat the relationship between speech waveforms and their transforms as transparent: given one, the other can be computed mechanically. Using the terminology of linguistics, different transforms, physical and mathematical, merely provide 'notational variants' of the waveform. But as every student of high school algebra knows, a clever change in notation, sometimes just a simple linear transformation of the variables, can make the difference between solving a problem or not. We will assume that transformations are freely available, but we will not assume that we know which one is the best. In particular, we will not assume that those yielding articulatory parameters are a priori better than those for which a straightforward physical interpretation is not readily available.

Let us collect speech waveforms in a set K of real-valued real functions. K is a proper subset of the set of all such functions: all waveforms are continuous, differentiable, and have many other properties that distinguish them as corresponding to speech rather than to music or other non-speech noises. Unfortunately, currently we do not fully understand what these properties are and we can not formulate a model that will generate all and only speech waveforms corresponding to phonologically well-formed utterances of a given language, dialect, or individual. Conversely, given an arbitrary waveform that meets some general criteria concerning amplitude (loudness) and the distribution of energy among various frequencies, we can not decide automatically whether this waveform corresponds to a well-formed utterance. The only way to get a grip on K is sampling it: we can record speech waveforms generated by human speakers. Sufficiently large samples will also provide information about the distribution of

the data, e.g. about the frequency of a particular form occurring in some context. Such information can be summarized by a probability measure M over K, implicitly conditioned by the sampling procedure (whether we use male of female speakers, adults or children, speakers of British or American English, speaking slowly or fast, informally or formally, and so on). The production task embodied in the sampling procedure can also be extended to, or verified by, a perception task, where listeners are asked to provide a transcription in some broad phonemic alphabet. This will associate to each waveform x a transcription $A(x)$, with the inverse A^{-1} of the transcription function A filling a role analogous to that of the *interpretation function* in Montague semantics.

To say that an utterance is composed of a succession of (micro)segments is to say that the transformed waveform at every instant can be described by interpolating between steady-state canonical targets. In order to create a mathematically well-defined problem of interpolation, we need to specify a class of functions among which we search for the optimal interpolation function. In the literature, this problem is hardly ever addressed explicitely: reading between the lines we can conclude that most authors have a class of functions in mind that only contains 'smooth' functions. For the sake of concreteness we assume that 'smoothness' means the piecewise continuity of low-order derivatives of the transform. We also need to define a figure of merit reflecting how closely the function approximates the target points or how closely two functions approximate one another. But before we turn to this problem in 3.2, let us briefly summarize the canonical target model for the concatenative case.

Given a set of transcribed utterances, we model the population from which this sample was drawn by specifying (i) an inventory of concatenative units P; (ii) a definition of these units in terms of constant (or piecewise constant) target values and the probability of their occurrence; (iii) a transformation T that maps each utterance into a slowly changing function; and (iv) a distance measure that tells us how well the concatenatively generated function in the model approximates the slowly changing function obtained by transforming a given waveform. Almost all synthesis and recognition models in the literature fit this general scheme, though the details of the implementation vary widely.

An important example of canonical target models is provided by parametric speech synthesizers such as MITalk (Allen et al 1987). For each unit (i), its acoustic parameters are initially looked up in a table, and are subsequently modified by parameters describing the speaker and the context. To the extent that these models are driven by underlying (or orthographic) representation, the probability of occurrence required in (ii) above has to be deduced from the language model used to drive the synthesis and from the distribution of speaker-dependent parameters – the latter is implicitly assumed to match the population of the speakers to be modeled. The typical transformation is linear predictive coding (LPC), and the typical distance measure is Euclidean (L_2) distance.

Hidden Markov Models (Baker 1975) can also be thought of as canonical tar-

get models, with the hidden states corresponding to the underlying concatenative units, and the output distributions defining the extensional models of these units. In modern systems, the underlying units are typically triphones (phonemic units further specified by their immediate left and right neighbors), the transformation is usually mel cepstral, and the distance is computed in terms of (log) probabilities. While in synthesis systems variability across individuals is only modeled implicitly and variability within the speech of a single individual is weakly modeled, recognition systems are very explicit about capturing both kinds of variation.

3.2 Partial targets and superposition

Trajectories corresponding to speech waveforms seldom show abrupt jumps. To the extent we find discontinuities, we find them mostly in the higher derivatives, and even there they are disguised and smoothed out by a great deal of noise. Since there are not that many discontinuities, models producing stepwise constant trajectories will not fit the data well. Indeed, the necessity of some *smoothing* procedure, roughly corresponding to phonetic/phonological assimilation processes, has long been recognized, and almost every model will have some smoothing built in. This results in a loss of direct coupling between segmental psychological units and their realizations, because the form a smoothed segment takes will depend on the neighboring segments. Since the neighbors themselves depend on their neighbors, some interaction between non-adjacent segments is predicted to be possible. At the phonetic level such interactions have been demonstrated by Öhman (1966). At the phonological level, non-local interactions have been known at least since Pāṇini 8.4.1-39, and they constitute the strongest grammatical evidence in favor of the model that takes segments to be bundles of distinctive features.

Though smoothing is often hidden in the details of the signal processing and engineers consider it a routine step of minor importance, for the phonologist it marks an important step away from direct models. Since smoothing is a separate computational procedure that uses the canonical targets as input, we can no longer say that psychological units are directly related to waveforms. As the smoothed trajectory is no longer piecewise constant, the input constants can no longer be directly recovered. The floodgates are open: once smoothing is permitted there is nothing to stop us from using more elaborate computational procedures and view the relationship between the psychological units and their eventual realization as less direct than type/token relationship. To be sure, speech engineers would, for the most part, deny that the pre-smoothed stage constitutes a theoretically significant intermediate level of representation. But from the perspective of the phonologist the question needs to be asked differently: is there anything special about assimilation that makes a dedicated computational step necessary? From the available phonological evidence, the answer is neg-

ative. Assimilation is just a form of spreading, no different in character from other phonological regularities such as vowel harmony.

Following this argument to its logical conclusion means that we must employ the same abstract specification method for local and for long-range phenomena. This method will be called *superposition* because it is completely analogous to the superposition method long familiar from physics: instead of computing the effects of various forces one by one, we first compute the combined force and only then compute its effect. The role of physical forces will be played by phonological constraints between psychological elements and their realizations, and the role of the physical quantity (action) to be minimized will be played by utterance distance.

According to the Cherry-Halle-Jakobson program, features F_i should be directly equated to regions a_i of the acoustic space so that if a segment p can be characterized as $[+F_1, -F_2, +F_3]$ it must be in the part of the acoustic space that is given by $a_1 \cap \overline{a_2} \cap a_3$. While the simplicity of this *intersective* model (formalized in greater detail in Kornai 1993) is appealing, the failure of the program to yield workable acoustic models suggests that a more complex approach explicitely incorporating feature interaction is required. We will model a feature F by the probability distribution of the corresponding canonical target ϕ, but do not assume that the targets will partition the acoustic space. Finding the acoustic realization x of a phoneme p composed of features F_i will require solving a minimum distance problem: if the distance is error squared, this will be simply the center of gravity of the intended targets ϕ_i. Since the targets are not points but probability distributions, the solution itself will be a probability distribution. The intersective model could also be extended probabilistically, but it would define the solution by multiplying the relevant probabilities, while in the present model the solutions are the result of a more complex additive procedure (a form of convolution).

As an illustrative example, consider two features F_1 and F_2 represented by one-dimensional distributions ϕ_1 and ϕ_2. Let ϕ_1 have a narrow peak of probability .99 at $x = 0$ and another peak of probability .01 at $x = 1$, and let ϕ_2 have .01 probability at $x = 0$ and .99 at $x = 1$. With two such incompatible features, the probability of a phone x satisfying both is $<.02$, and the only values that x can take are 0 and 1, if we use the multiplicative model. However, if we use the superposition model, the requirement to satisfy F_1 and F_2 means that we select the point x minimizing the error $(x-0)^2 + (x-1)^2$, which is at $x = 0.5$ with probability $>.98$. Since the distributions ϕ_i can not be measured directly the model has considerable freedom inasmuch as 98% of the data ($x = 0.5$) could be just as well described by starting from ϕ_1' and ϕ_2' which have their peaks not at 0 and 1 but at -0.5 and +1.5 or at any other points equidistant from 0.5. But different data sets will impose different constraints and if the number of features remains below the number of segment types the overall system will still have a unique best solution.

To see that the ϕ_i can not be directly measured it is sufficient to consider a relatively simple pair of features, tonal H and L, which have a single, well understood, and relatively easily measurable acoustic correlate, the fundamental frequency F_0. The distribution ϕ_H of F_0 in tone bearing units phonologically marked H will be nearly identical to the distribution ϕ_L obtained for those marked L, even for a single speaker, as long as the language exhibits significant downdrift. It is true that phrase-initial H and L would be statistically distinguishable, but unless we know that there is such a thing as downdrift we do not have any reason to inspect the phrase-initial portion of the data separately. If we are aware of downdrift, we are in the business of fitting more complex statistical models (involving e.g. exponential decay to a baseline), so the point that the relevant parameters can not be directly measured but must be computed is evident. The simplest model incorporating downdrift would be one where it is not F_0 but its derivative $F_0^{'}$ which is computed by interpolation to two constant targets, a positive one corresponding to H and a negative one corresponding to L. Needless to say, the mathematically simplest model is not necessarily the best one, but the larger point that a choice between models can be made on a purely statistical basis remains valid quite independent of whether it is fruitful to consider derivatives in the trajectories (Furui 1986).

The central idea of the superposition model is that the same process, interpolation, applies not only across successive elements, but also across simultaneous elements. Smoothing requires interpolation between successive points, and modeling features requires finding the minimum error trajectory among several possibly conflicting targets. The basic requirements of the model remain the same: (i) an inventory of nonconcatenative units F; (ii) a definition of these units in terms of constant (or piecewise constant) target values and the probability of their occurrence; (iii) a transformation T that maps each utterance into a slowly changing function; and (iv) a distance measure that tells us how well the succession of multiple partial targets generated in the model approximates the slowly changing function obtained by transforming a given waveform.

Combinatorial theories of phonology give us a good handle on the inventory (i). Traditional acoustic phonetics along the lines of Peterson and Barney provides a great deal of information about (ii) in the segmental case, and this naturally extends to cases such as tridirectional vowel features where the features can be realized in isolation. Speech engineering research offers a variety of transforms (iii) and solves the problem of finding the appropriate distance measure (iv) by combing transformations until distance computations reduce to the mathematically simplest Euclidean distance. A good example is provided by Itakura-Saito distance which is intimately connected to the linear predictive coding (LPC) transformation (Atal and Hanauer 1971). The LPC transform associates an acoustic filter $T(x)$ to a sound, and the distance of x to another sound y is measured by inverse filtering y through $T(x)$ and computing the energy of the residual sound. Another example, more directly relevant to the quest

for a direct acoustic characterization of distinctive features, is the canonical distortion measure introduced by Baxter (1994), which is built on the basis of a known classification into discrete categories. When combined with Dynamic Time Warping (for an overview see Sankoff and Kruskal 1983) the same way as in Itakura (1975), this measure would yield zero distance between two waveforms if and only if they have the same phonemic transcription.

4 Conclusion

In this paper we have outlined a framework for analyzing, comparing, and refining analytic theories of phonology, and presented an alternative to the intersective model of features. While our concept of underlying representations as discrete cognitive structures generated from segmental or autosegmental units is fairly standard in phonology, our view of surface representations admitting only automatically measurable physical properties of the speech signal will perhaps be called 'phonetic' by some phonologists. But the goals of a theory are no doubt more important than the name we use to describe it, and the means we use to achieve the goal of a practicable theory are more important still. We proposed that phonology, if it is to produce results useful for scientists and engineers interested in speech, must go beyond the combinatorial enumeration of discrete structures and explicitely address the issue of variable physical realizations, their mathematical transformations, and provide a distance measure that tells us how well the model fits the data.

REFERENCES

Allen, Jonathan, M. Sharon Hunnicutt and Dennis Klatt (1987). *From text to speech: the MITalk system*. Cambridge University Press.

Anderberg, Michael R. (1973). *Cluster analysis for applications*. Academic Press, New York.

Anderson, John and Colin Ewen (1987). *Principles of dependency phonology*. Cambridge University Press, Cambridge.

Atal, Bishnu S. and Suzanne L. Hanauer (1971). Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America* **50** 2, 637–655.

Baker, J. (1975). The DRAGON system – an overview. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **ASSP-23** 24–29.

Baxter, Jonathan (1994). *The canonical metric for vector quantization*. Ms, Flinders University of South Australia.

Bell, Alexander M. (1867). *Visible speech*. Simpkin, Marshall, and Co,, London.

Bird, Steven (1990). *Constraint-based phonology*. PhD Thesis, University of Edinburgh.

Bird, Steven and T. Mark Ellison (1994). One-level phonology: autosegmental representations and rules as finite automata. *Computational Linguistics* **20** 1 , 55–90.

Booij, Geert and Jerzy Rubach (1987). Postcyclic versus postlexical rules in lexical phonology. *Linguistic Inquiry* **18** 1, 1–44.

Browman, Catherine P. and Louis Goldstein (1990). Tiers in articulatory phonology with some implications for casual speech. In *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, John Kingston and Mary E. Beckman, (eds.) Cambridge University Press, 341–376.

Cherry, Colin (1956). Roman Jakobson's distinctive features as the normal coordinates of a language. In *For Roman Jakobson*, Morris Halle, (ed.) Mouton, The Hague.

Cherry, Colin, Morris Halle and Roman Jakobson (1953). Toward the logical description of languages in their phonemic aspect. *Language* **29** 34–46.

Chomsky, Noam (1956). Three models for the description of language. *I. R. E. Transactions on Information Theory* **IT-2**.

Chomsky, Noam and Morris Halle (1968). *The sound pattern of English*. Harper & Row, New York.

Cole, Ronald A., Michael S. Phillips, Robert Brennan, Ben Chigier, Rich Green, Robert Weide and Janet Weaver (1986). Status of the C-MU phonetic classification system. In *Proc DARPA Speech Recognition Workshop*. Palo Alto, CA, 1–5.

Cole, Ronald A., Richard M. Stern, Michael S. Phillips, Scott M. Brill, Andrew P. Pilant and Philippe Specker (1983). Feature-based speaker-independent recognition of isolated English letters. In *ICASSP-83*. Boston, MA, 731–733.

Everitt, Brian (1980). *Cluster analysis*. Halsted Press, New York.

Fant, Gunnar (1973). *Speech sounds and features*. MIT Press, Cambridge.

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.

Furui, Sadaoki (1986). Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **ASSP-34** 1, 52–59.

Goldsmith, John A. (1976). *Autosegmental phonology*. PhD Thesis, MIT.

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

Halle, Morris and Jean-Roger Vergnaud (1987). *An essay on stress*. MIT Press, Cambridge.

Hulst, Harry van der (1989). Atoms of segmental structure: components, gestures, and dependency. *Phonology* **6** 2, 253–284.

Itakura, Fumitada (1975). Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **ASSP-23** 67–72.

Jakobson, Roman (1939). Observations sur le classment phonologique des consonnes. In *Proc. 3rd International Congress of Phonetic Sciences*. 34–41.

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

Kaplan, Ronald and Martin Kay (1994). Regular models of phonological rule systems. *Computational Linguistics* **20** 3, 331–378.

Kaye, Jonathan, Jean Lowenstamm and Jean-Roger Vergnaud (1985). The internal structure of phonological elements: a theory of charm and government. *Phonology (Yearbook)* **2** 305–28.

Keating, Patricia (1985). CV phonology, experimental phonetics, and coarticulation. *UCLA Working Papers in Phonetics* **62** 1–13.

Kingston, John and Mary E. Beckman, (eds.). *Papers in laboratory phonology I: between the grammar and physics of speech*. Cambridge University Press.

Kiparsky, Paul (1982). Lexical morphology and phonology. In *Linguistics in the morning calm*, I.-S. Yang, (ed.) Hanshin, Seoul, 3–91.

Klatt, Dennis H. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America* **82** 3, 737–793.

Kornai, András (1993). The generative power of feature geometry. *Annals of Mathematics and Artificial Intelligence* **8** 37–46.

Kornai, András (1994). Relating phonetic and phonological categories. In *Language Computations*, Eric Sven Ristad, (ed.) DIMACS 17, American Mathematical Society, 21–35.

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.

Liberman, Alvin M., F. S. Cooper, D. Shankweiler and M. Studdert-Kennedy (1967). Perception of the speech code. *Psychological Review* **74** 431–461.

Mohanan, K. P. (1982). Lexical phonology, PhD Thesis, MIT.

O'Shaughnessy, Douglas (1987). *Speech communication, human and machine*. Addison Wesley.

Öhman, Sven E. G. (1966). Coarticulation in VCV utterances: spectographic measurements. *Journal of the Acoustical Society of America* **39** 151–168.

Paradis, Carole (1988). On constraints and repair strategies. *The Linguistic Review* **6** 71–97.

Peterson, G. E. and H. L. Barney (1952). Control methods used in the study of vowels. *Journal of the Acoustical Society of America* **24** 175–184.

Prince, Alan and Paul Smolensky (1993). *Optimality theory. Constraint interaction in generative grammar*. Ms, Rutgers University and University of Colorado.

Remez, R. (1979). Adaptation of the category boundary between speech and nonspeech: a case against feature detectors.. *Cognitive Psychology* **11** 38–57.

Repp, B. (1983). Categorial perception: issues, methods, findings. In *Speech and Language: Advances in basic research and practice*, N. Lass, (ed.) vol. 10, Academic Press, New York.

Rissanen, Jorma (1978). Modeling by the shortest data description. *Automatica* **14** 465–471.

(1983). In *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison* , Sankoff, David and Joseph B. Kruskal, (eds.) Addison-Wesley, Reading, MA.

Scobbie, James (1993). Constraint violation and conflict from the prespective of declarative phonology. *Canadian Journal of Linguistics* **38** 2.

Scobbie, James M. (1991). *Attribute value phonology*. PhD Thesis, University of Edinburgh.

Stevens, Kenneth (1972). The quantal nature of speech: evidence from articulatory-acoustic data. In *Human communication: a unified view*, E. David Jr and P. Denes, (eds.) McGraw-Hill, New York.

Stevens, Kenneth N. and Sheila E. Blumstein (1981). The search for invariant acoustic correlates of phonetic features. In *Perspectives on the study of speech*, P. Eimas and J. Miller, (eds.) Lawrence Erlebaum Associates, Hillsdale, NJ.

Watrous, Raymond L. (1991). Current status of Peterson-Barney vowel formant data. *Journal of the Acoustical Society of America* **89** 5, 2458–2459.

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