DIMACS Series in Discrete Mathematics
and Theoretical Computer Science
Volume , 0000

Relating phonetic and phonological categories

ANDRÁS KORNAI

April 5, 1993

. Recent proposals to treat phonetic representations as the semantic interpretation of phonological representations are technically problematic to implement. The main difficulty is that phonological representations are discrete while phonetic representations are continuous which makes the standard method of describing semantic interpretation as a homomorphism between sortally equivalent algebras hard to generalize. The paper solves the technical problem by introducing a notion of homomorphisms. First, the operation of concatenation is defined in the usual way for strings and as 'continuation' for continuous scalar-vector functions with finite support, and the set of phonetic representations is equipped with a measure. Next a.e. homomorphisms are rigorously defined and the semantic relationship between phonological and phonetic categories is made explicit in terms of these homomorphisms. Finally constant target (triphone) models, which play a central role in speech recognition, are reconstructed in this semantic framework.

0. Introduction

From a cognitive standpoint human speech can be described as a succession of linearly and hierarchically organized discrete units including *sounds, syllables,* and *words*. From a physical standpoint, it can be described like any other sound, by plotting air pressure as a continuous function of time. Finding the exact relationship between this *acoustic waveform* and the cognitive units is a task of immense practical significance: reliable automatic speech recognition and synthesis algorithms would revolutionize information technology and would greatly

1991 . Primary 68S05, 68T10, 92K20; Secondary 03B65, 28B10, 28C.

The author was supported in part by a seed grant from the Center for the Study of Language and Information, Stanford University. Thanks to Gábor Elek (Purdue University), John Coleman (AT&T Bell Labs), and the conference participants and anonymous reviewers for comments and criticisms.

This paper is in final form and no version of it will be submitted for publication elsewhere.

aid the handicapped. In the current academic division of labor, the cognitive aspects of speech are studied under the heading of *phonology,* while the physical aspects, including the biological mechanisms employed in speech perception and production, are studied under the heading of *phonetics.* The formal apparatus of these two fields reflects their primary concerns: phonologists tend to employ graphs, automata, rewrite rules, and other tools of discrete mathematics, while phoneticians prefer Fourier analysis, differential equations, and other tools of continuous mathematics. The goal of this paper is to develop a class of mathematical models that can bridge the gap between the two by effectively specifying the relationship between the discrete phonological categories and the continuous phonetic observables.

There is a growing consensus in linguistics (Pierrehumbert 1990, Bird 1990, Coleman and Local 1991) that the relationship between the cognitive units and their phonetic realization is structurally analogous to the relationship between symbols and their meaning. This suggests that in order to understand the phonology/phonetics relationship better, we should bring the technical tools of semantics to bear. There is a rich tradition of formal semantics, starting with the work of Russell and Frege at the turn of the century, that we can draw on. In the mathematical domain, where the intended meanings are relatively simple, this tradition can be considered definitive since the work of Tarski (1949) and Carnap (1947). In the linguistic domain, where the intended meanings are far more elusive, no definitive formal semantics has yet emerged, but a host of important technical contributions were made, with Montague's "Universal Grammar" (1970) serving as the foundation of most subsequent work. The central idea behind these developments is that the relationship between syntax and semantics is *compositional* and thus should be captured with the aid of a homomorphism between algebras of the same sort. Unfortunately, the algebra of phonological structures is a discrete, finitely generated structure, while the set of phonetic realizations has a continuous structure which is differential geometrical rather than combinatorial in nature. This apparent sortal incompatibility puts considerable technical obstacles in the course of developing a model that treats phonetics semantically. The main contribution of this paper is in showing how these obstacles might be overcome by suitably weakening the notion of homomorphism.

The paper assumes a certain mathematical sophistication on the part of the reader, but no familiarity with phonology or phonetics – a brief survey of these fields will be provided in Section 1. The key technical innovation designed to deal with the problem of incompatibility is introduced in Section 2, where the notion of *almost everywhere* homomorphism is rigorously defined. In Section 3 the resulting formal theory is applied to the special case of *constant target* (triphone) models which play a central role in computationally inspired theories of phonetics.

1. Phonetics and phonology

The traditional division of labor between phonetics and phonology is embodied in the following “speech chain”:

(1.1)

phonological representation
Generative Rules
articulatory representation
Acoustic Synthesis
speech waveform
Peripheral Processing
auditory representation
Central Processing
phonological representation

As a first approximation, *phonological representations* can be conceived of as linear strings of meaningful units such as phrases or words, endowed with *constituent structure,* conceptualized as a planar tree in which the nodes correspond to the units, the edges correspond to the “constituent of” relation, and the linear order of the daughter nodes corresponds to the linear order of the constituents of the mother node. Such trees are usually presented in a linearized notation using brackets or boundary symbols to denote the constituent breaks. In the course of analyzing the units into ever smaller constituent parts, meaningful units soon give way to purely phonological units, such as *feet, syllables*, and *segments,* whose justification is to be found in the regularities of the sound system, rather than in the constraints imposed upon the language by its syntax or semantics (see Nespor and Vogel 1986).

While in general there is excellent correspondence between the higher units established on the basis of meaning (e.g. syntactic phrases or words) and units of roughly the same size established on the basis of phonological criteria (e.g. phonological phrases or words), this correspondence degrades as the units get smaller: in fact between the minimal meaningful units or *morphemes* and the minimum default pronunciation units or *syllables* there is no real correspondence just a vague overall tendency for morphemic and syllabic breaks to coincide. Accordingly, the need for “readjustment rules” mediating between grammatical and phonological constituent structures has long been recognized (see Bierwisch 1966), and tree structures depicting purely phonological constituency (including a distinguished daughter constituent, the *head* or *most prominent* constituent) are routinely used (see Hayes 1980). A different notational system for expressing prominence is the *metrical grid* originating in the work of Liberman (1975), and later elaborated in Liberman and Prince (1977), Selkirk (1984) – a syncretic formalism is presented in Halle and Vergnaud (1987).

It is fair to say that the issues of phrasing, rhythm, stress, intonation, traditionally grouped together under the heading of *suprasegmental* phenomena, are not nearly as well understood by phonologists as issues of *segmental* or *subsegmental* phonology. The basic insight, that segments are composed of smaller, temporally parallel units called *distinctive features* predates generative and even structuralist phonology, but the systematic development of the idea is due to Jakobson (for an overview, see chapter 5 of Anderson 1985), and the definitive formalization in terms of "feature matrices" (actually, vectors) is given in Chomsky and Halle (1968). At the risk of considerable oversimplification, the idea of feature decomposition can be said to rest on the observation that production of a minimal speech segment or *phoneme* involves the coordinated activities of several articulators such as the lips, the tongue blade, the tongue body, and so on. Subsequent developments in generative phonology, in particular the advent of *autosegmental* phonology are largely aimed at preserving this basic insight while removing the constraint known as *absolute slicing* which requires the articulators to act in absolute synchrony. For the linguistic motivation of autosegmental phonology see Goldsmith (1990), and for a formal analysis of feature structures and "geometries" see Kornai (in press).

Returning to (1.1) above we can now see how phonology begins with a discrete structure (string, tree, or more complex graph) and ends with an *articulatory representation* something like a musical score, with the "orchestra" being the human vocal tract, the "instruments" being the independently controllable articulatory organs, and the "notes" being the positions these organs can assume. However, it is important to keep in mind that the *gestural score* provided by phonological theory is in no way comparable in precision to Western musical notation. First, articulator positions are given in grossly simplified and idealized physiological terms such as lip rounding vs. no lip rounding, high tone vs. low tone. Second, the absolute and relative timing of the gestures leading into and out of the prescribed positions is also simplified and idealized. Third, and most important, these gestural scores do not come with a precise set of interpretative conventions. There are no absolute statements like a "high tone is 300Hz" or even relative statements such as "a long vowel is twice as long as a short one".

Given a set of measurements describing how the absolute dimensions of a speaker's vocal tract change in time, acoustic phonetics is to a remarkable extent able to synthesize a corresponding speech waveform, though as O'Shaughnessy (1987:312) notes: "Practical implementations of such a vocoder have yet to be found, due to our limited understanding of how to accurately model the relationship between vocal tract parameters and the speech spectrum, particularly for excitation within the tract". However, even if our understanding of articulatory synthesis advanced to the point of perfection, we would still need to deal with the information gap between the output of generative phonology and the input required for acoustic modeling.

The classical work of Liberman et al. (1959) culminating in Klatt's MITalk (see Allen et al. 1987), and the more modern autosegmental synthesis models such as Browman and Goldstein (1985, 1989), Fujimura (this volume) all rely extensively on the proper setting of various continuous parameters describing the physical dimensions of the vocal tract and the (absolute and relative) timing of articulatory gestures. Some of these parameters, such as acoustic tube length or overall speech rate can be directly manipulated to describe different voice qualities and speech styles. Others, such as fundamental frequency, need to be controlled by complex models that take not only physiological but also language- and dialect-particular and even strictly grammatical factors into account. Yet others appear as solutions to various equations describing constraints on the overall parameter space.

The question how the grammatical and the extragrammatical, the physiologically determined and the consciously controllable parameters interact is far from resolved. Although parametric synthesizers are quite capable of mimicking (adult male) voices, finding the appropriate parameters to drive such systems is a formidable task (Holmes 1983). As a practical matter, the highest quality speech is synthesized by algorithms that bypass the acoustic synthesis stage entirely, working with samples of pre-recorded natural speech instead. The most successful speech recognition algorithms are also based on direct acoustic pattern matching (Baker 1975, Klatt 1980). However, such systems lose sight of the basic cause and effect model in (1.1) and because they sacrifice parametric control they can provide no theoretical insight into the factors that contribute to the variability of the speech signal.

As we progress further along the speech chain, the situation becomes progressively worse. Only the most optimistic hypothesis about peripheral processing, the *motor theory* of Liberman *et al.* (1967) promises that we can gain as much understanding of perception as we have of production – every other theoretical model is severely constrained by the limitations of our ability to trace nerve impulse patterns back to the central nervous system. Unfortunately, no system of auditory representations has ever been proposed that would match even the limited detail offered by articulatory representations. The motor theory simply fills this void by equating perceptual and articulatory categories.

Finally, for want of an empirically testable alternative, theories of central processing generally assume that speech recognition is simply the converse of speech synthesis. Since the rules of generative phonology are context-sensitive rules permitting deletion, analysis by synthesis is computationally intractable. While this problem might be remedied by constraint-based theories of phonology (Wheeler 1981, Koskenniemi 1983, Bird and Klein 1990, Scobbie 1991) or by strictly limiting the variety of rewrite rules available for generation (Archangeli and Pulleyblank in press), analysis by synthesis algorithms still have to rely on rules of synthesis, and these generally bypass the articulatory/perceptual stage and go directly from the phonological representation to the acoustic waveform.

To summarize this discussion, both phonology and phonetics have amassed a large body of data and created very sophisticated theories concerning their respective domains. Phonologists are largely able to generate discrete gestural specifications from discrete underlying (cognitive) representations, and in principle, though not in practice, their model is neutral between analysis and synthesis. Phoneticians are largely able to generate continuous waveforms, or analytically equivalent continuous representations (such as spectra and cepstra, see e.g. Flanagan 1972, Rabiner and Schaefer 1979) from continuous multivariate descriptions of the vocal tract and the excitation source, and given sufficient information about one or the other, the acoustic model is also reversible. However, there are no clear-cut interpretative principles relating the gestural scores to fully specified physical descriptions of the vocal tract, and there is no effectively computable theory of *categorial perception* i.e. the emergence of discrete perceptual units from continuous input. Therefore, a theory of relating the output of one to the input of the other is still missing, and the use of articulatory/auditory representations, while convenient for establishing the boundaries of these disciplines, is by itself unable to resolve the sortal incompatibility problem.

In Section 2 we present the basic ideas of a formal theory capable of systematically relating the discrete structures used in phonology to the continuous structures used in phonetics. How these general ideas can be applied for specific varieties of phonetic and phonological theory will be discussed in Section 3.

2. The basic model

Our formal model of human speech is built on the set of acoustic waveforms that can be produced by speakers of a given language: these form a fixed subset K of the real-valued real functions T_1.[1] The phonetic/phonological structure that K is endowed with will be captured as an ordered triple (M, A, P) where M is a *probability measure* over K (in other words, a σ-additive function from certain sets of K to to the real interval [0,1]), P is a finite set of symbols $p_1, ..., p$ and A is a mapping from K to P (the free monoid generated by P). Intuitively, M reflects the probability of a given set of waveforms being produced, P reflects the segment inventory of the language in question, and A is the assignment of a string of segments (phonemic transcription) to a given waveform. In what follows these intuitive ideas will be gradually replaced by rigorous definitions that can serve as the basis for investigating the problem in an analytic setting.

2.1. The statistical structure. Since certain classes of waveforms occur more frequently than others, K comes equipped with a statistical structure, which will be captured with the aid of a measure M. The following equation expresses an important aspect of the relationship between the measure M and the domain of the mapping A, namely that *noises* (acoustic waveforms with no

[1] Or the set of scalar-vector functions . When the dimension of the range space is irrelevant, the subscript is suppressed.

phonemic interpretation) have zero probability. Elements of T outside K form a set of zero measure:

$$M(\{x \in T : x \notin dom(A)\}) = 0 \tag{2.1}$$

It needs to be emphasized that M is *not* the standard Lebesgue measure on T. Part of our goal will be to *represent* M using Lebesgue-Stieltjes measures of the appropriate sort, but we must wait until Section 3 to define exactly what we mean by such a representation.

2.2. The phonological structure. Let us now turn to the set P of segmental (phonemic) symbols. The phonological description of the language provides not only P (a simple list of the elements that appear in the phonemic inventory) but also a *feature analysis* based on a small, presumably universally fixed, list of features $F_1, ..., F$. We can take the feature analysis to be a family of mappings from P to $G_1, ..., G$ where the G are finite sets (typically of cardinality one or two) that describe the possible values of the feature F . In order to avoid having to speak of partial structures, each G is defined as containing the symbol U(nderspecified) so that instead of a family of mappings we can talk of a single mapping

$$f : P \rightarrow \Pi\ _{=1}G \tag{2.2}$$

As we shall see in 3.1, the issue of defining A by "pulling back" over f is intimately related to the issue of *invariant clues* for features (cf. Stevens and Blumstein 1981). But for the moment, let us ignore the issue of feature decomposition altogether and concentrate on the free monoid P . By defining A as a direct mapping with domain K and range P we avoid the complexities of distinguishing peripheral and central processing from one another.

2.3. The phonetics-phonology homomorphism. The relationship between continuous phonetic and discrete phonological categories is captured by the function A which maps elements of K onto elements of P . The key idea of the whole formalization is that both the domain and the range of A comes naturally equipped with an operation of concatenation, and A preserves this operation *almost everywhere.*

In order to elucidate the natural concatenation operation on K we will restrict our attention to a subset of the full set T_1 (or T). Those functions to $\mathbf{R}$ (or $\mathbf{R}$) that have finite support will be called *curves.* Since the phonemic content of a curve g is independent of the time it is uttered, we have

$$A(gSt) = A(g) \tag{2.3}$$

for every $g \in K, t \in \mathbf{R}$, where the *shift* of g by t, gSt is defined by $(gSt)(x) = g(x+t)$. For every $g \in K$ that has support $[a, b)$ we define the *left translate* Lg of g to be $gS-b$ and the *right translate* Rg of g to be $gS-a$. Now the concatenation gh of two curves g and h is defined as the curve $LgRh$. It is trivial to verify that concatenation is indeed associative (mod S).

The notion of concatenation can be lifted in the usual fashion from concatenation of curves to concatenation of a curve g and a set of curves $H \subset K$ by defining

$$gH = \{gh : h \in H\} \tag{2.4}$$

Similarly to the concatenation of two sets $G, H \in K$ will be defined as

$$GH = \{gh : g \in G, h \in H\} \tag{2.5}$$

Now we can define precisely what we mean by A being a homomorphism "almost everywhere". If $G \subset K$ such that for almost every $g \in G, A(g) = p$, and $H \subset K$ such that for almost every $h \in H, A(h) = q$,

$$M(\{x \in GH : A(x) \neq pq\}) = 0 \tag{2.6}$$

3. Representation

If a structure is not fully understood, we can still gain insight into its properties by *representation theory* i.e. by studying the homomorphic images of the structure in some other, better understood structures. In the previous section we have defined the structure (M, A, P) on K, and here we turn to the issue of what it means to *represent* this structure in simpler euclidean structures where coordinates correspond to physically measurable quantities.

We will create representations in three steps. In 3.1 we define subsets of curves that correspond to phonemes or archiphonemes (autosegmental feature-combinations) and show how P can be represented in terms of *cardinal targets*. In 3.2 we define what it means to represent the probability measure M. Finally, in 3.3 we complete our sketch of phonological representation theory by representing A in terms of *interpolation* between stationary targets.

The phonological and phonetic theories used in this process serve only to illustrate how the broad semantic framework outlined in Section 2. can be used to specify the relationship between phonetic and phonological categories. There is no presumption that these are the *correct* phonological or phonetic categories: they were chosen because they are still rather widely used in applied phonology and phonetics, and because using more complex (and in all likeness more correct) phonological and phonetic theories would require the formalization of many assumptions that have no bearing on the main point.

3.1. Phonemes and archiphonemes. The key idea behind this formalization is Frege's insight that the interpretation of the whole must be derivable by simple, uniform means from the interpretation of the parts. Formally, this idea of *compositionality* can be captured by Montague's method of requiring that the interpretation mapping be a homomorphism: whenever we create a complex structure from two or more constituents, the meaning of this structure must also be composed of the meaning of the constituents. Let us now see how this idea applies for phonology.

As a first step of let us investigate those sets of curves that are the inverse images of the generators of P . Since A is invariant under translation, we will concentrate on the sets $K \quad = R(A \;\, ^1(p\;))$ and in general $K \quad = R(A \;\, ^1(\alpha))$. In an idealized model, where no phonological or phonetic *assimilation* takes place, we would simply have $K \quad = K \;\, K$ for every α and $\beta \in P$. Notice that this leaves the problem of segmentation open: we know that for every curve c such that $A(c) = \alpha\beta$ there are curves a and b such that $c = ab$ and $a \in K$, $b \in K$ but we do not know how to find such an a and b.

In a considerably less idealized model, where assimilation of adjacent segments is permitted, we can introduce *triphones* K as follows. For the sake of convenience we enlarge the phonemic inventory P with a new symbol p_0 that will conceptually correspond to silence (unfilled pause) and concentrate on curves c such that $A(c) = p_0 p \;\; p \;\; p \;\; ...p \;\; p_0$. For these the triphone hypothesis guarantees the existence of curves $b \;\; , b \;\; , ..., b$ such that (taking $i_0 = i \;\, _{+1} = 0$) $c = b \;\; b \;\; ...b$ and the following equations hold:

$$b \quad \in K \tag{3.1}$$

In a model that will permit unbounded assimilation within the limits of autosegmental association domains, the inventory of representative elements is even more complex. So far we have dealt with the monoid-homomorphism $A : K \rightarrow P$ that mapped curves to phonemic transcriptions. The phonology of the language also provides a finite set of *features* $F_1, ...F$ with corresponding *value sets* $G_1, ..., G$ as well as a set of mappings $f \;\; : P \rightarrow G$ together defining the *feature chart.*

$$\begin{array}{ccc} K & = & K \\ A & & A \\ P* & \longrightarrow & G \\ \pi & & \pi \\ P & \longrightarrow & G \end{array} \tag{3.2}$$

Since the f can be naturally lifted from $P \rightarrow G$ to $P \quad \rightarrow G$ mappings, we can always combine the resulting f with A to yield $A \quad = f \;\; A$. Conceptually this corresponds to direct transcription of curves into sequences of feature vectors in the manner of Jespersen (1904), Chomsky and Halle (1968). In the truly autosegmental case absolute slicing fails, so instead of A we only have the A in (3.2). Thus the pivotal elements in the representation will have to be constructed by means of intersection from the inverse images of the strings in G ($1 \leq j \leq s$). Of particular importance are the inverse images of U under some A : these contain those curves which are *underspecified* for F in their entirety. Phonetically this can mean two different things: either the relevant feature is *undefined* for the curve in question (e.g. tone for voiceless stops) or it is defined but its value is freely chosen, e.g. by considerations of articulatory inertia.

Finally, note that in cases of more sophisticated feature geometries involving class nodes (Clements 1985), the direct process of taking inverse images must be replaced by an indirect process of descending to the terminal nodes of the geometry recursively. However, this descent is over trees that are subtrees of some fixed finite template, and will therefore always terminate in a fixed number of steps. This means that the process of composition is more complex, but the basic idea of compositionality is still valid.

3.2. Lebesgue-Stieltjes representation. The inverse images collected so far have a great deal of phonetic similarity: the curves in any single set receive the same phonemic transcription. Thus it is reasonable to assume that they are all variants of the same ideal curve, which we will call the *cardinal* curve. At least for steady-state phonemes, this curve is in some sense constant. To make this idea more precise we introduce the notion of *transformation*. In practice a transformation is some mapping B from T_1 to T that is defined locally (usually over a 100ms or even shorter window) but in principle we could consider any B that maps waveforms onto the trajectory of a single point in some euclidean *feature space*. For example, by means of a short-term Fourier transformation we can turn the original swiftly oscillating one-dimensional curve into a constant, or at least very slowly changing spectrum which in turn can be characterized by n slowly changing parameters such as the first n formants or cepstral coefficients[2].

However, representing elements of P by the averages of their inverse images is only part of the task: in order to represent the full structure (M, A, P) on K we must also find a means of representing M. This goal is achieved if we find a transformation B such that the probability measure M is transformed by B to some kind of *natural* probability density function. It is not obvious whether the most natural density for this purpose is uniform, as suggested by vector quantization techniques, gaussian, as suggested by continuous density Hidden Markov Models, or some other density. The present formalization can remain largely neutral on this issue: let us simply assume some natural density function D : $\mathbf{R} \rightarrow \mathbf{R}$, where the Q are sets of parameters (such as means

[2] In practice it is advantageous to include the derivatives of cepstral coefficients as well, in which case "constant" means "constant in phase space".

and covariances) depending on the p . Given some set $K\ \subset K$ composed of constant curves of length l we know that all temporal cross-sections $B(K\ )(l\ )$ are the same for $0 \leq l\ < l$ – let us denote this time-invariant cross-section by $B(K\ )$. For the probability measure M to be represented by the euclidean feature space by the density functions D the following must hold:

$$(3.3) \qquad M(K\ |l) = \ _{(\quad)}\ D\quad d\lambda$$

where λ is the ordinary euclidean Riemann-Lebesgue measure. The right-hand side of this expression is independent of l. This is made possible by relying on the assumption that for each $p \in P$ there is a *duration density* μ such that the probability of a set of curves L having cross-section K and any length between l and $l + \Delta l$ is

$$(3.4) \qquad M(L\ ) = \ ^{+\Delta}\ M(K\ |t)\mu\ (t)dt$$

Notice that the duration densities employed here are tied to the linear units distinguished. If we assume an invariant syllable or segment-concatenation model each syllable or segment will have a characteristic duration density function. If we assume local assimilation, each triphone will have its own duration density, and so on.

3.3. Interpolation. So far we considered only steady-state segments that can be characterized in terms of a constant target. But what happens when the curve only approximates the target, or if it oscillates around the target extremely rapidly? In the former case, it can still be a lower probability version of the same segment, while, in the latter case, it is more likely to be a nonspeech noise of some sort. This contrast shows that in general distance between two curves can not be defined as the integral of pointwise distances not even for curves with the same support.

Given the physical nature of the vocal tract, it makes sense to value curves which arise as the result of some smooth interpolation between cardinal targets more highly than others. To make this idea more precise would require the specification of some functional, such as average curvature, to be minimized, perhaps in combination with some penalty incurred when target points are only approximated but not reached. Let us concentrate on the case when each feature is linked independently to the root tier. Ideally, each dimension of the feature space corresponds to one feature F and each member of the value set G represents a constant target on that axis, with U(nderspecified) denoting either a lack of value or any form of smooth interpolation. Thus in a representation the mappings A are replaced by the components B of the transform B in such a manner that for each curve $g \in K$ and each feature F the j-th component of the transformed curve $B\ (g)$ reaches the targets corresponding to $A\ (g)$ in the appropriate order, with synchrony among the various components defined by the pattern of association lines and the components jointly solving the minimum problem for the smoothness functional. In diagram:

(3.5)
$$\begin{tikzcd} T \arrow[r, hook] & G \\ B & \mathrm{CarT} \\ P & H \arrow[l] \end{tikzcd}$$

where the $CarT$ -s are assignments of cardinal targets to feature values, and S is the solution[3] corresponding to the specified string of cardinal targets.

To make this more concrete, let us pick (admittedly arbitrarily) the continuity of the first three derivates as our smoothness condition, with piecewise cubic polynomials as the class of functions used for optimization. For an independently linked tier, such as the tonal tier, we first need to specify a feature (say F_9) and its value set G_9: let us say G_9 contains the values Low, Mid, and High (plus U). Next we need to specify a transform B_9 that will compute from each waveform $g(t)$ a "tonal projection" $B_9(g)(t)$: in this particular case we actually know how to effect such a transformation by *pitch tracking*. Now, if a given waveform g has segmental projection $A(g) = s_1 s_2 .... s$ and tonal projection $A_9(g) = T_1 T_2 .... T$ (with one-to-one association, for the sake of simplicity) and we have cardinal targets (pitch values) CarT$_9$(L) = 400, CarT$_9$(M) = 500, CarT$_9$(H) = 700[4] the task becomes one of finding a piecewise cubic with continuous first three derivates that take on the appropriate cardinal values at $t_1 < t_2 < .... < t$.

As is well known, there is no unique solution to the above problem: rather, we have a 2-parameter solution for each set of "knots" $t_1 < t_2 < .... < t$ or, since the location of the knots is not fully known, essentially a k + 2-parameter family of solutions.[5] The inverse image (under $B_9{}^{1}$) of the space of solutions gives a constraint on the set of original curves K , and the other tiers (plus their linking patterns) provide other constraints. In a detailed parametric representation the sets K are recoverable analytically and the probability measure of various subsets can be expressed in terms of the distribution of cardinal target values. However, it should be kept in mind that the parametric description of these distributions must include not only grammatical, dialectal, and social factors, but also the physiological characteristics and individual style of the speakers.

[3] More likely a set of solutions: for this a powerset mapping should be added to the diagram.

[4] Since the aim is not to present a theory of tone but to illustrate the man features of the model, phenomena specific to tone, such as downdrift, depressor consonants, etc. are ignored here. The numerical values are chosen for children rather than adult males – see below.

[5] In fact, the structure is even more complex, for if some of the was underspecified in the strong sense that it can not bear tone the curve is split into parts, while underspecification in the weaker sense of tone being present but not distinctive reduces the number of knots.

4. Conclusions, further directions

One of the biggest problems for models incorporating the scheme presented in (1.1) is segmentation. As Glass and Zue (1988) show, finding an exhaustive, non-overlapping partitioning of the timeline into subintervals representing the temporal extent of the segments is still a major problem. In the model presented here, segmentation becomes less of an obstacle to a precise statement of the problem, since only target points are considered and even these need not coincide for different tiers.

The distance measures developed by speech engineers interested in a practical solution to the recognition problem are specifically designed to be invariant under a broad class of time warping functions relating two n-dimensional curves of possibly different length (Gray and Markel 1976). The present approach suggests invariance under an even broader class of warps, one where the n featural dimensions can be subject to independent warping as long as the association structure is not violated.

To conclude, the problem of specifying the phonetics-phonology mapping is an important practical problem that has so far been attacked largely by directly exploiting the statistical structure of K via M, as in Hidden Markov Modeling (Baker 1975), or by indirectly exploiting its differential geometrical structure via some articulatory transform B. The semantically inspired formalism presented here suggests a more abstract approach that puts the emphasis on the topological structure of K: the phonemic transcription associated to a waveform by A is to be viewed as a topological invariant of the curve.

The continuous/discrete dichotomy in the focus of this paper might even turn out to be epiphenomenal, as argued by Browman and Goldstein (1990), who view the discrete phonological units as emergent from the nonlinear dynamics of the articulatory system. However, in order to make this a verifiable claim, a large number of parameters must be explicitly specified, together with their range of variation. Specifying the appropriate topology should be the first step towards the realization of the more ambitious goals of specifying the appropriate metric and measure.

References

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

Anderson, Stephen R. 1985 *Phonology in the Twentieth Century: Theories of Rules and Theories of Representations* University of Chicago Press, Chicago

Archangeli, Diana and Douglas Pulleyblank 1993 *The content and structure of phonological representations* MIT Press, Cambridge MA (in press)

Baker, James K. 1990 *Stochastic modeling for automatic speech understanding,* Readings in speech recognition (Alex Waibel and Kai-Fu Lee, eds) Morgan

Kaufmann, San Mateo CA, pp. 297-307

Bierwisch, Manfred 1966 *Regeln für die Intonation deutscher Sätze* Studia Grammatica **7** 99-201

Bird, Steven 1990 *Constraint-Based Phonology* PhD Thesis, University of Edinburgh

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

Browman, Catherine P. and Louis Goldstein 1985 *Dynamic modeling of phonetic structure,* Phonetic Linguistics (Victoria Fromkin, ed) Academic Press, New York pp. 35-53

Browman, Catherine P. and Louis Goldstein 1989 *Articulatory gestures as phonological units* Phonology **6** 201-251

Browman, Catherine P. and Louis Goldstein 1990 *Representation and reality: physical systems and phonological structure* Journal of Phonetics **18** 411-424

Carnap, Rudolf 1947 *Meaning and necessity* University of Chicago Press, Chicago

Clements, George N. 1985 *The geometry of phonological features* Phonology Yearbook **2** 225-252

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

Coleman, John and John Local 1991 *The "No Crossing Constraint" in Autosegmental Phonology* Linguistics and Philosophy **14** 295-338

Flanagan, James 1972 *Speech Analysis, Synthesis and Perception* Springer Verlag, New York

Glass, James R. and Victor W. Zue 1988 *Multi-level acoustic segmentation of continuous speech* Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-88), New York, pp. 429-432

Goldsmith, John A. 1990 *Autosegmental and metrical phonology* Basil Blackwell, Cambridge MA

Gray, Augustine H. and John D. Markel 1976 *Distance measures for speech processing* IEEE Acoustics Speech and Signal Processing **24** 380-391

Halle, Morris and Jean-Roger Verganud 1987 *An essay on stress* MIT Press, Cambrdige MA

Hayes, Bruce 1980 *A metrical theory of stress rules* PhD Thesis, MIT, Cambridge MA

Holmes, J. N. 1983 *Formant Synthetizers: Cascade or Parallel* Speech Communication **2** 251-273

Jespersen, Otto 1904 *Lehrbuch der Phonetik* B.G. Teubner, Leipzig

Klatt, Dennis H. 1980 *SCRIBER and LAFS: two new approaches to speech analy-*

sis, Trends in Speech Recognition (Wayne Lea, ed) Prentice-Hall, pp. 529-555

Kohonen, T. 1988 *The "neural" phonetic typewriter* IEEE Computer **21** 11-22

Kornai, András 1993 *The generative power of feature geometry* Annals of Mathematics and Artificial Intelligence (in press)

Koskenniemi, Kimmo 1983 *Two-level Morphology: a general computational model for word-form recognition and production* Helsinki Department of General Linguistics, University of Helsinki Publication **11**

Liberman, Alvin M., Frances Ingemann, Leigh Lisker, Pierre Delattre, and F. S. Cooper 1959 *Minimal rules for synthesizing speech* Journal of the Acoustic Society of America **31** 1490-1499

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

Liberman, Mark 1975 *The Intonational System of English* PhD Thesis, MIT, Cambridge MA

Liberman, Mark and Alan Prince 1977 *On stress and linguistic rhythm* Linguistic Inquiry **8** 249-336

Montague, Richard 1970 *Universal Grammar,* Formal Philosophy (Richmond Thomason, ed) Yale University Press, New Haven CT pp. 222-246

Nespor, Marina and Irene Vogel 1986 *Prosodic phonology* Foris, Dordrecht

O'Shaughnessy, Douglas 1987 *Speech Communication, Human and Machine* Addison Wesley, Reading MA

Pierrehumbert, Janet 1990 *Phonological and phonetic representation* Journal of Phonetics **18** 375-394

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

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

Selkirk, Elisabeth O. 1984 *Phonology and Syntax: The Relation Between Sound and Structure* MIT Press, Cambridge MA

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

Tarski, Alfred 1949 *Introduction to logic and to the methodology of deductive sciences* Oxford University Press, New York

Wheeler, Deirdre W. 1981 *Aspects of a categorial theory of phonology* PhD dissertation, University of Massachusetts, Amherst

: kornai@csli.stanford.edu