Comparing Prosody Across Many Languages

Fred Cummins, Felix Gers, Jürgen Schmidhuber
Istituto Dalle Molle di Studie sull'Intelligenza Artificiale
Corso Elvezia 36
CH 6900 Lugano
Switzerland
{*fred,felix,juergen*}*@idsia.ch*

I.D.S.I.A. Technical Report IDSIA-07-99

Abstract

An empirical metric is developed to measure similarity or distance between the prosodic systems of diverse languages. Two physical variables which carry prosodic information, F_0 and amplitude envelope modulation, are used as inputs in a pairwise language discrimination task. A novel neural network model is trained to discriminate among pairs of languages drawn from a set of ten languages in a telephone speech database. Network performance is used to obtain a distance metric for each pair, allowing overall similarity relationships within the 10 language set to be identified. Results are presented for F_0 alone, in which Mandarin and Japanese are found to be clearly distinguished from all other languages in the database, and French is found to be clearly distinguishable, even from related Indo-European languages. Using amplitude envelope modulation alone, a cluster of languages usually described as syllable-timed is identified. Vietnamese and Farsi are found to be rhythmically distinct from languages conventionally described as stress-, syllable- or mora-timed. When both variables are provided to the networks, the results provide an empirical measure of prosodic distance among the languages.

PACS numbers: 43.70.Kv 43.72.Bs

1 Introduction

Languages differ. But the range of variation found across languages is not unlimited. By studying similarity and difference across languages, we may hope to delimit the range of possible linguistic variation, which is a central goal of linguistic inquiry. The simplest case arises when we have exhaustive knowledge of *a priori* possibilities, as, for example, in identifying the canonical order of subject, object and verb, which must be some permutation of the set {S,O,V}. When the underlying primitives are harder to identify, and more numerous, the situation quickly becomes more complex. Phonemic inventories are easier to compare across languages than syllable types, in part because segmental theory has reached a more mature stage than the phonological theory of suprasegmentals.

This paper deals with two aspects of prosody, which we loosely call speech rhythm and intonation. More precisely, we examine the ways in which F_0 modulation and amplitude envelope modulation vary across languages. It is probably fair to say that the study of intonational systems is less well developed than the study of segmental properties, while the study of speech rhythm remains in its infancy. In order to effectively compare languages based on these variables, a method must be found which can provide an empirical metric of similarity without recourse to still unknown phonological primitives.

Fundamental frequency is the physical variable which encodes lexical tone and phrasal intonation. It is an important marker of accent or stress in every language. In addition, speaking style, affect, rate and gender all contribute to observed F_0 contours. Because several phonological primitives, and a host of extra-linguistic factors may affect F_0 contours simultaneously, it seems unlikely that we could propose or defend a simple taxonomy of language based only on observed F_0 patterns.

Some principled ways in which languages make use of tone and intonation are known. We can identify languages which employ lexical tone, such as Mandarin Chinese or Zulu, those which use lexically based pitch accents, like Swedish or Japanese, and stress accent languages, such as English or German. This crude categorization obscures significant differences within each category; for example, Swedish has two contrasting tones, and minimal pairs which differ only in tone exist. Japanese, on the other hand, contrasts only tone and no tone. The task of identifying an inventory of phonological primitives for the intonational system of any given language is difficult, and a theoretical concordance which would allow comparison of any two arbitrary languages simply does not exist.

There is some evidence that raw F_0 contours may serve to discriminate among languages. Itahashi, Zhou and Tanaka (1994) found that they could generate a parametric representation of the F_0 contours extracted from speakers of 6 languages (3 Asian, 3 European), and that discriminant functions could then be found which separated all 6 languages. It is difficult to infer the utility of their representation due to the small amount of data used, and the absence of any test of the discriminant procedure on independent material. Some researchers attempting to develop systems for automatic language identification (LID) have found F_0 to be of limited use in improving overall identification performance, but currently, LID is best performed using segmental and phonotactic information (Hazen and Zue, 1994).

The most direct study of the utility of prosodic features in LID is summarized in Thymé-Gobbel and Hutchins (1996). Discrimination between pairs of English, Japanese, Mandarin and Spanish samples was attempted using 224 individual features and feature pairs. F_0-based features included several representations of pitch contour and pitch difference across syllables. Amplitude modulation was specified by the amplitude contour on a syllable and the difference in maximum amplitude across syllables. The low frequency FFT of the amplitude envelope was included, as was the syllable rate. Many durational features were also computed. Discrimination was done using log-likelihood ratio functions based on histograms of the various features or feature pairs. Overall, F_0-based features proved to be the most effective, either alone or in combination with almost every other feature. Amplitude-based features performed very poorly, yielding results no better than than chance in most cases. The authors noted that "familiarity with the linguistic specifics of a language allows us to make predictions about the usefulness of particular prosodic features in the discrimination of a given language-pair", and they concluded, rather pessimistically, that using amplitude features is a very poor LID strategy".

While we are still far from a mature theory on intonation that could be applied to all langauges, phonological approaches to speech rhythm are even less well formulated. To some, rhythm refers to the succession of strong and weak units (syllables, accents), and their hierarchical groupings (Liberman and Prince, 1977; Kiparsky and Youmans, 1989). To others, rhythm refers to the perception of regular intervals in speech. The latter approach, formulated first by Pike (1945) and elevated to the level of dogma by Abercrombie (1967), led to the postulation of a simple 3-class taxonomy of stress-, syllable-, and, more recently, mora-timed languages. English, Spanish and Japanese are often considered to be canonical examples of these respective classes. A host of empirical studies has failed to provide much evidence for a language taxonomy based on equal time intervals. In particular, it is not the case that stressed syllable onsets are evenly spaced in English (Classé, 1939; Dauer, 1983), nor are syllables of equal duration in Spanish (Pointon, 1980; Wenk and Wioland, 1982; Dauer, 1983). The hypothesis that Japanese morae are spoken with equal duration has received somewhat more support in Port, Dalby and O'Dell (1987), where it was demonstrated that the average mora duration remains approximately constant as the number of mora in a word increases.

The lack of support for the simple 3-way rhythmic classification has in no way prevented it from serving as a model for studies of other languages. The conclusion most often reached, however, is that language X is not easily classified as stress- or syllable-timed (less work has been done on mora timing). Thus "Tamil can be called neither stress-timed nor syllable-timed" (Balasubramanian, 1980), "Korean speech rhythm is in between the two opposite types" (Lee et al., 1994), "Not all Kannada listeners perceive any rhythm...it is not clear what is the actual phonetic basis of the perceived rhythm" (Savithri, 1995), "[Brazilian] Portuguese is in the process of changing from a syllable-timed language to a stress-timed language" (Major, 1981). Similarly inconclusive results were obtained by Dauer (1983) for English, Spanish, Italian, Greek and Thai.

The absence of empirical support for a rhythmic taxonomy based on isochrony contrasts with the finding that some languages are perceived as rhythmically similar. For example, Nazzi et al. (1998) demonstrated that French newborns could discriminate low-pass filtered English and Japanese sentences, but not English and Dutch ones. Similarly, they showed that English and Dutch (both stress-timed) could be collectively discriminated from Spanish and Italian (both syllable-timed), but English and Italian could not be discriminated from Spanish and Dutch. In this and related studies, speech is lowpass filtered (typically with a cut off frequency of 400 Hz), in an attempt to reduce segmental information, while leaving prosodic information intact. Both F_0 and amplitude envelope modulation (at least) remain together in the resulting stimuli, and it is not possible to assess the role each of these variables may play. Nazzi et al. (1998) use the term *rhythmic distance* to capture the fact that languages may be more or less prosodically similar, without necessarily endorsing a grouping into discrete and well-defined classes.

The comparison of intonation and rhythm across languages is further complicated by the large effect of speaking style. Prosodic patterns observed in carefully enunciated isolated sentences may have little in common with those of conversational speech. We are acutely aware of the difference between the vigorous extemporaneous public speaker and one who reads a prepared paper in soporific monotone. But which should serve as representative of "the language"?

Strong evidence for rhythmic constraints on the intervals between the onsets of stressed syllables in English has been demonstrated in a task specifically designed to elicit rhythmic speech (speech cycling) (Cummins and Port, 1998). When subjects repeat a phrase like "beg for a dime" in time with a perfectly regular metronome, there are three, and only three, stable patterns, each of which corresponds to a simple 2- or 3-beat rhythm, based on the onsets of accented (stressed) syllables. Similar tendencies have since been observed in Japanese (Tajima, 1998), although the languages differ in the relative stability and prominence of the underlying units (syllables/morae, stresses/accents). Although speech cycling studies may serve to uncover differences in rhythmic patterns across languages, the strong constraints observed are specific to the phrase repetition task employed.

Ramus and Mehler (1999) examined the ability of native French speakers to discriminate between English- and Japanese-like stimuli, where the stimuli were constructed by resynthesizing original utterances with all consonants and vowels replaced by archetypes. In the two most relevant conditions, the resulting stimuli were designed to preserve only intonation (all segments replaced by /a/, original F_0 contour preserved) or rhythm (consonants replaced by /s/, vowels by /a/, flat intonational contour). They found good discrimination could be obtained using the latter, rhythmic, stimuli, but the former, F_0-based stimuli did not support discrimination.

In what follows, we present a comparison among 10 languages, in which F_0 and amplitude envelope modulation are extracted from unscripted telephone speech. One or both of these inputs are fed to a neural network trained to discriminate between a pair of languages. The network employed is a novel recurrent architecture which has demonstrated the ability to solve difficult problems in which relevant information is distributed across several time scales.

We adopt the working hypothesis that networks will be more successful in discriminating among dissimilar languages, and that a measure of network success in the discrimination task can thus form the basis for a metric of similarity among languages. By combining the results from all possible pairwise discrimination tasks, we can establish an overview of similarity relations within the 10-language set as a whole, based on F_0, amplitude envelope modulation and on both together. In this way, we hope to compare prosodic systems, without recourse to any specific set of hypothetical phonological primitives. We seek to establish an empirical measure of similarity which does not presuppose discrete classes, thus putting flesh to the concept of prosodic, or rhythmic, distance (Nazzi et al., 1998).

The present work is an extension of a study in which we trained networks to discriminate among the languages English, German, Spanish, Mandarin and Japanese (Cummins et al., submitted). Our principle findings there were that both amplitude envelope modulation and F_0 could serve to discriminate at better than chance levels among languages which were typologically distinct. Thus, English, German and Spanish were readily distinguishable from Mandarin and Japanese based on F_0 alone, but were indistinguishable among one another. Surprisingly, Mandarin and Japanese were not well discriminated based on F_0, but were quite well discriminated based on amplitude envelope modulation.

2 Speech materials

For all experiments we use speech from the Oregon Graduate Institute Telephone Speech corpus (OGI_TS) (Muthusamy et al., 1992), which is available from the Linguistic Data Consortium. This database contains recordings of telephone conversations in which speakers responded in their native language to a series of prompts designed to elicit speech on a continuum from specific lexical items to unconstrained monologue. The database includes approximately 200 speakers for each of 11 languages. Several groups have published results based on this corpus, making comparison across systems feasible (Hazen and Zue, 1994; Zissman, 1995; Thymé-Gobbel and Hutchins, 1996).

The OGI corpus provides recordings acquired over commercial telephone lines of speakers responses to an automatically generated series of prompts. The prompts elicit lexically constrained responses (e.g. "Please recite the seven days of the week"), topic-specific but otherwise unconstrained responses (e.g. "Tell us something that you like about your home town") as well as about 60 seconds of completely unconstrained speech per caller. Recordings were made at 8000 Hz with 14-bit resolution. The OGI_TS database was established with the specific goal of developing LID systems, and the data are thus noisy real world speech, in which dialects within a language are freely mixed, hesitations and speech errors are frequent, and channel noise is not negligible.

As training data, we use only the responses to the four topic-specific questions, each of which provides up to 10 seconds of speech. This yields four recordings per speaker, and we use 90 speakers per language. Data were available for the following 10 languages (Hindi is not included due to incomplete data): English, Farsi (Persian), French, Japanese, German, Korean, Mandarin Chinese, Spanish, Tamil and Vietnamese. This set includes representatives of the principle intonational (Mandarin, Japanese, German), and rhythmic types (Spanish, Japanese, English), as well as languages whose affiliations are less clear. Of these languages, much less phonetic and phonological analysis has been done on the set {Farsi,Vietnamese,Tamil} than on the remaining languages. Five belong to the large Indo-

European family (English/German: Germanic; Spanish/French: Italic; Farsi: Indo-Iranian), Tamil is a Dravidian language, Vietnamese is arguably Austro-Asiatic, Mandarin is Sino-Tibetan, and Korean and Japanese are both isolates, though each have been linked at times to the Altaic family. The set of languages thus is diverse, without claiming to be comprehensive.

2.1 F_0 preprocessing

F_0 traces were obtained using the pitch_track algorithm (Cambridge University Engineering Department), which provides a smooth pitch estimate for every 1 ms of speech, or 0 for frames deemed unvoiced. This estimate was first differenced, downsampled to 100 Hz using a sliding rectangular window, smoothed using a simple 15 point rectangular window average, and rescaled to lie within the range [-1,1]. We refer to this input as ΔF_0.

2.2 Amplitude envelope modulation

The amplitude envelope was computed by filtering the speech with a low-order Butterworth bandpass filter centered at 1000 Hz with a bandwidth of 500 Hz. This effectively removes F_0 and fricative energy, leaving formant information. Results of both Cummins (1997), and Scott (1993), suggest that the amplitude variation in this frequency range is important for the perceived rhythm of speech. The filtered speech was rectified by taking absolute values of each sample and then smoothed using a Butterworth lowpass filter with cut-off at 10 Hz. It was then first-differenced, downsampled to 100 Hz, re-smoothed and rescaled in the same manner and using the same parameters as for ΔF_0. This input will be called ΔEnv.

Network inputs thus consist of a time series of two real numbers, ΔF_0 and ΔEnv. Figure 1 illustrates inputs for one speech file (English, ca. 4 sec duration).

3 Long Short-Term Memory

In a "conventional" recurrent neural network, hidden units receive input from an external source, as well as from other hidden units (and possibly output units); their outputs are a non-linear function of their weighted net input. As a result, the state of the network at any given time depends not only on present input, but also on previous states. In this way, the network state can be viewed as a short term memory, as opposed to the long term memory instantiated in the (slowly changing) weights between units. Standard gradient-based training methods, such as Back Propagation Through Time (BPTT) and Real-Time Recurrent Learning (RTRL), propagate error back in time in order to properly apportion credit among the network weights. As pointed out in Hochreiter (1991), these models suffer from the limitation that the magnitude of the error signal either decays exponentially or blows up, leading to instability, as more and more history is taken into account. This makes it difficult or impossible for these networks to learn to solve problems which require storage of information over long time lags. Any attempt to counter this by simply sampling the input more coarsely in time loses the opportunity of attending to fine temporal structure. The result is an ability to attend to information which is distributed across a range of time scales.

An attempt to overcome these limitations is the Long Short-Term Memory (LSTM) model, first presented in Hochreiter and Schmidhuber (1997). In this model, the hidden units of a conventional recurrent neural network are replaced by memory *blocks*, each of which contains

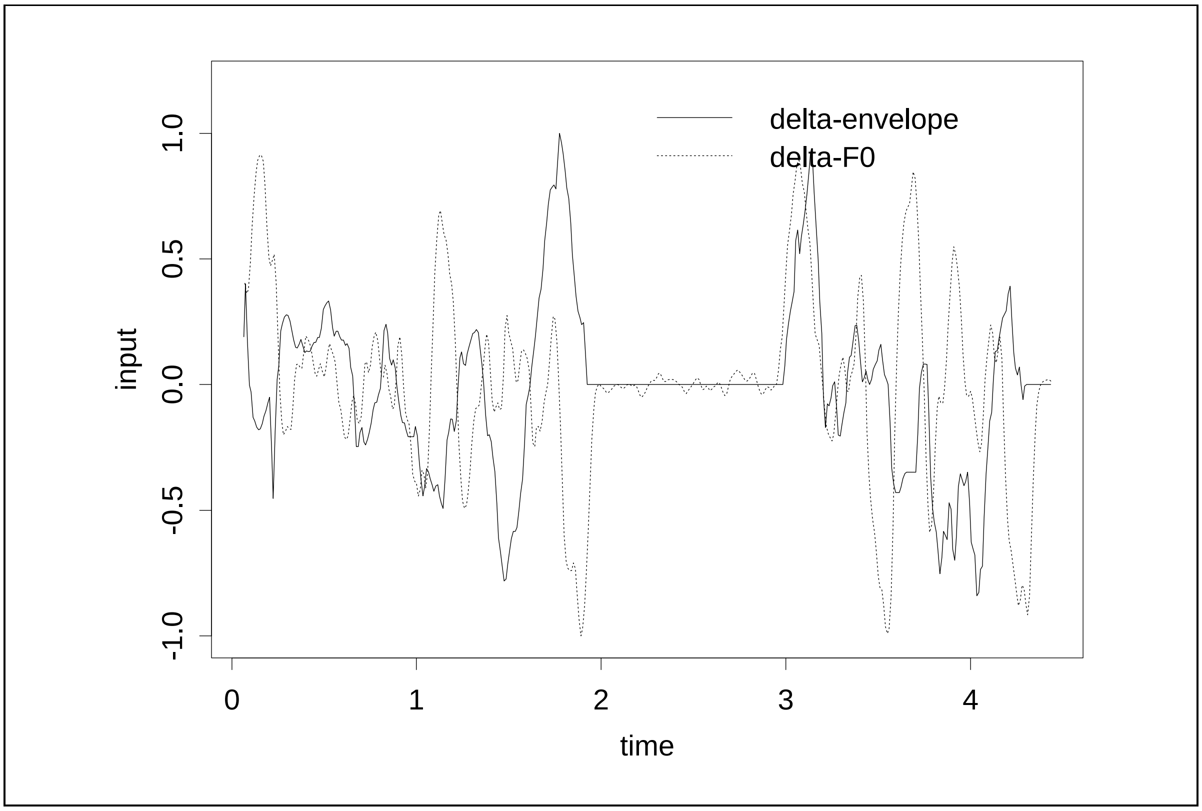

Figure 1: Network inputs for one English sound file. Solid line is the modulation of the band-limited amplitude envelope, ΔEnv, dashed line is change in F_0, ΔF_0.

one or more memory *cells*, illustrated in Figure 2. At the heart of the cell is a simple linear unit with a single self-recurrent connection with weight set to 1.0. In the absence of any other input, this connection serves to preserve the cell's current state from one moment to the next. In addition to the self-recurrent connection, cells receive input from input units and other cells (see below).

While the cells are responsible for maintaining information over long periods of time, the responsibility for deciding what information to store, and when to apply that information lies with an input and output gating unit, respectively.

The input to the cell is passed through a non-linear squashing function ($g(x)$, typically the logistic function, scaled to lie within [-2,2]), and the result is then multiplied by the output of the input gating unit. Only if the input gate is sufficiently active does the signal enter into the cell. Similarly, nothing emerges from the cell unless the output gate is active. As the internal cell state is maintained in a linear unit, its activation range is unbounded, and so the cell output is again squashed when it is released ($h(x)$, typical range [-1,1]). The gates themselves are nothing more than conventional units with sigmoidal activation functions ranging over [0,1], and they each receive input from the network input units and from other cells and gates.

This division of responsibility—the input gates decide what to store, the cell stores information, and the output gate decides when that information is to be applied—has the effect that salient events can be remembered over arbitrarily long periods of time. Equipped with several such memory blocks, the network can effectively attend to events at multiple time

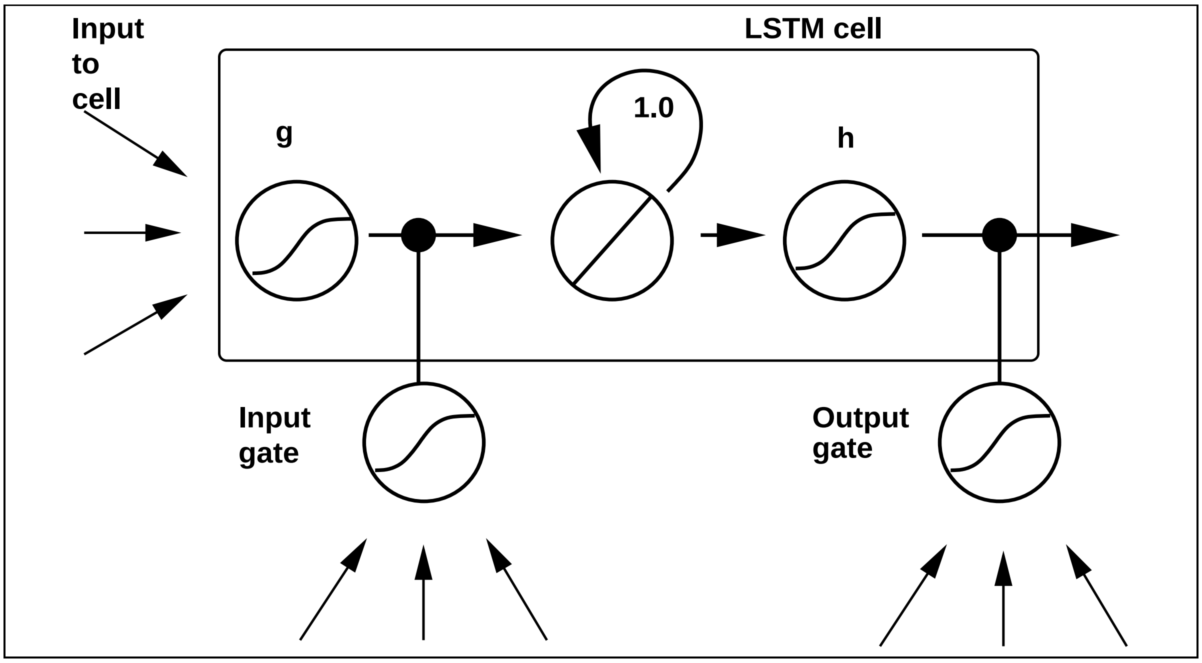

Figure 2: A single one-cell memory block. The block comprises a cell and two gating units. Cell input is squashed by $g(x)$, gated by an input gate and passed to the linear unit at the center. This unit's activation is the cell state. In the absence of any other input, cell state is preserved through the self-recurrent connection. Cell output is computed by squashing the cell's state ($h(x)$) and gating it with an output gate.

scales.

Network training is done using a combination of RTRL and a truncated version of BPTT. Full details are provided in Hochreiter and Schmidhuber (1997). The authors also present results from a series of experiments on both old and new problems, in which the effectiveness of LSTM in solving problems involving long-term temporal dependencies is demonstrated. The method can successfully solve problems in which information must be stored over time lags in excess of several thousand time steps, while established recurrent neural networks typically fail once time intervals of more than about 10 time steps must be bridged.

In all experiments reported herein, we used networks having one or two inputs (ΔF_0 and/or ΔEnv, as described above), three memory blocks each with a single cell, and a single output unit, whose activation indicates the network's decision in the discrimination task. Weights were provided from input to cells, gates and output unit, and from cells to cells, gates and output. No weights originating at the gates or at the output unit were provided. Biases, implemented as an additional input unit with constant value of 1.0, were provided for gates and the output unit, but not for cells. Weights and biases (except as detailed below) were randomly initialized at the start of each trial from a uniform distribution ranging over $[-0.2, 0.2]$. Gates on the first memory block had biases of -1, on the second, -2 and on the third, -3. This ramp-like negative biasing of the gates ensures that the activation of all gates is close to zero at the start of training, so that most learning occurs in the direct connections from input to output. As learning progresses, the biases become progressively less negative, and, one after another, the memory blocks begin to contribute to the network computation.

This is one approach to the general goal in network training of ensuring that separate hidden units or elements attempt to solve separate parts of the problem (Mozer and Smolensky, 1989).

Network parameters were initially fixed based on pilot experiments in which networks were trained to discriminate between English and Mandarin on a subset of the data later employed in the experiments reported here. We then tested the present approach on a 5 language subset from the OGI_TS database (English, German, Spanish, Mandarin, Japanese). Results have been presented in Cummins et al., (submitted). Based on our observations in that study, we selected the training regime outlined below. Networks in the present study are not trained for as long a period as in Cummins et al., (submitted) (50 vs 120 epochs), and we here use slightly smaller training and test sets. As the present study examined 45 pairwise comparisons among languages, under three different input conditions, with 10 networks trained on each comparison and input, extensive optimization of training or architectural parameters was not feasible.

3.1 Network training

In each network trial, the network was trained on 180 speech files from each of two languages. Initial network weights were assigned randomly from the interval [-0.2,0.2]. Speech data was input at the 100 Hz resolution described above. A language-dependent target of either 0 or 1 was provided throughout the entire sequence. Weight updates were made at sequence end, and the network state was reset after each sequence. A sequence was judged to have been classified correctly if the average output for the last 50 inputs (0.5 sec) was on the correct side of 0.5. At the end of each training epoch of 360 speech files, network performance was evaluated using a validation set of 180 independent files, half from each language. No learning was carried out during validation.

Training continued for 50 epochs with a learning rate of 0.002. The order in which sequences were presented was randomized at the start of each epoch. Weights were stored when performance on the validation set was optimal, and the final network evaluation consisted of presentation of a novel set of 180 speech files. In this manner, generalization performance is entirely independent of both the training and validation sets. Random guessing (chance performance) would lead to an average of 50% correct.

4 Results

4.1 Comparisons based on ΔEnv

For each possible pairwise combination of the 10 languages, 10 networks were trained with ΔEnv alone as input, for a total of 45 pairwise comparisons. Table I lists the average percent correct and standard deviation for each pairwise comparison. Discrimination at or above the level of 60% correct is indicated by figures in bold font. Performance ranges from 72.2% (Farsi-Japanese) to 50.7% correct (English-German). As with our previous study, we observe that Mandarin and Japanese are well discriminated on the basis of ΔEnv.

While information about individual pairwise comparisons can be read off from Table I, it is not easy to see groupings of languages which appear similar to the networks. To this end,

	Ge	Sp	Ma	Ja	Fa	Ko	Ta	Vi	Fr
En	50.7 (2.3)	55.0 (4.1)	52.0 (2.0)	50.3 (1.1)	**62.2 (3.2)**	54.3 (2.2)	**60.9 (6.1)**	55.1 (3.5)	59.9 (5.5)
Ge	–	52.8 (1.8)	52.3 (4.2)	50.3 (1.1)	**62.5 (2.7)**	56.5 (3.8)	53.0 (1.7)	**61.2 (4.9)**	56.7 (2.7)
Sp		–	52.7 (2.8)	**60.7 (1.6)**	56.6 (2.5)	51.3 (1.7)	54.0 (4.7)	56.5 (1.5)	50.3 (0.7)
Ma			–	**70.2 (1.1)**	52.5 (1.1)	52.3 (3.7)	52.6 (1.0)	54.7 (2.0)	50.7 (1.2)
Ja				–	**72.2 (1.5)**	**62.0 (6.7)**	55.7 (2.4)	51.5 (2.0)	59.7 (2.8)
Fa					–	57.8 (1.5)	59.0 (1.7)	**70.9 (2.1)**	58.3 (2.8)
Ko						–	50.1 (1.1)	59.0 (3.9)	53.8 (3.2)
Ta							–	53.6 (2.7)	56.0 (5.6)
Vi								–	**61.6 (4.6)**

Table 1: Average percent correct (and s.d.) for pairwise comparisons based only on ΔEnv.

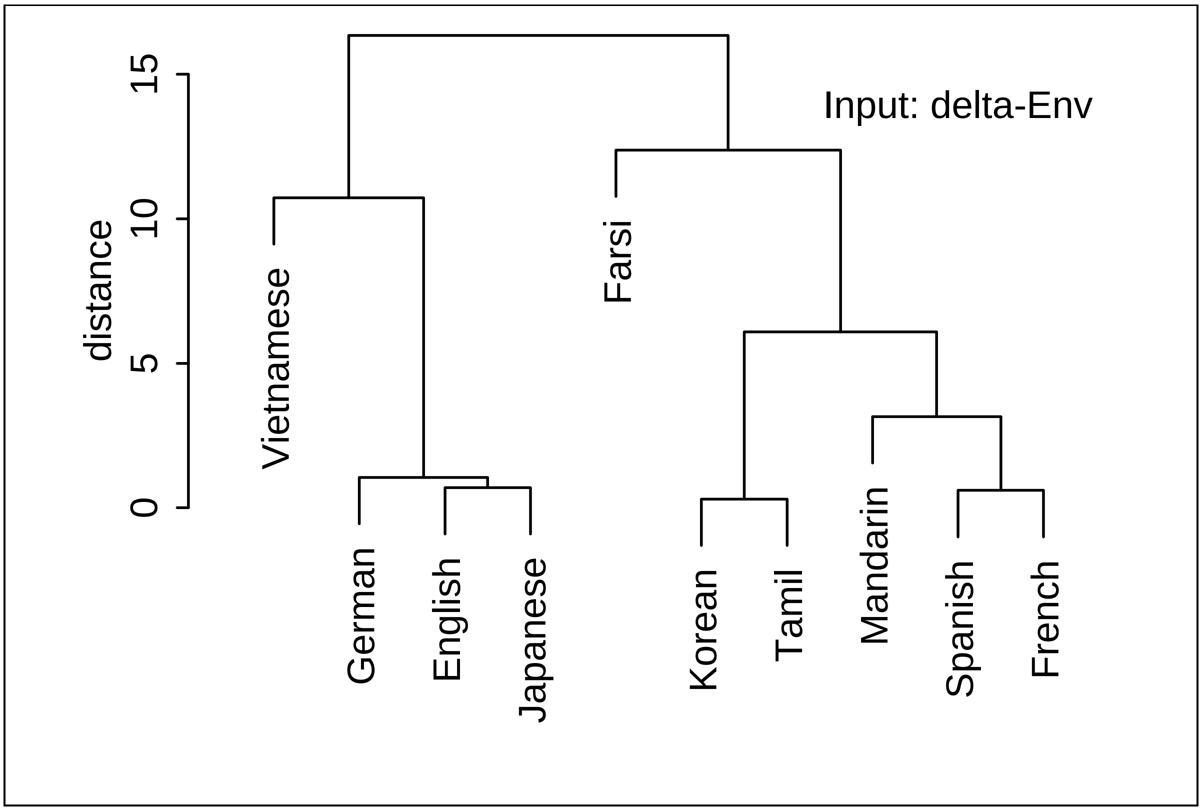

Figure 3:

an iterative clustering algorithm (hclust, method = "average", (Statistical Sciences, 1995)) was applied. This algorithm starts with each object in a group of its own. At each iteration, it merges the two closest groups into a single group, terminating when only a single group remains. The distance between two languages is $90 - d$, where d is the mean number of misclassified sequences, with $d = 90$ representing chance performance. The distance between two groups is the average of the distance between the points in one group and the points in the other. The resulting clustering tree is given as Figure 3.

In interpreting the cluster tree, it is important to bear in mind that distances between individual languages are obtained from average performance in the pairwise discrimination task. Distances between groups and a language and between groups and groups are based on group averages. Thus, in Figure 3, it is not possible to estimate the distance of Farsi from any of the other languages directly, for which purpose the reader is referred to Table I.

Of particular note in this figure is the first set of relatively tight clusters formed above the terminal elements. The three languages for which syllable timing has often been claimed, Spanish, French and Mandarin, do indeed group together, while English and German group together. Surprisingly, Japanese groups together with English and German. Korean and Tamil, both of which have been claimed to be neither syllable-timed nor stress-timed (Balasubramanian, 1980; Lee et al., 1994) form a distinct group, while Farsi and Vietnamese appear as outliers.

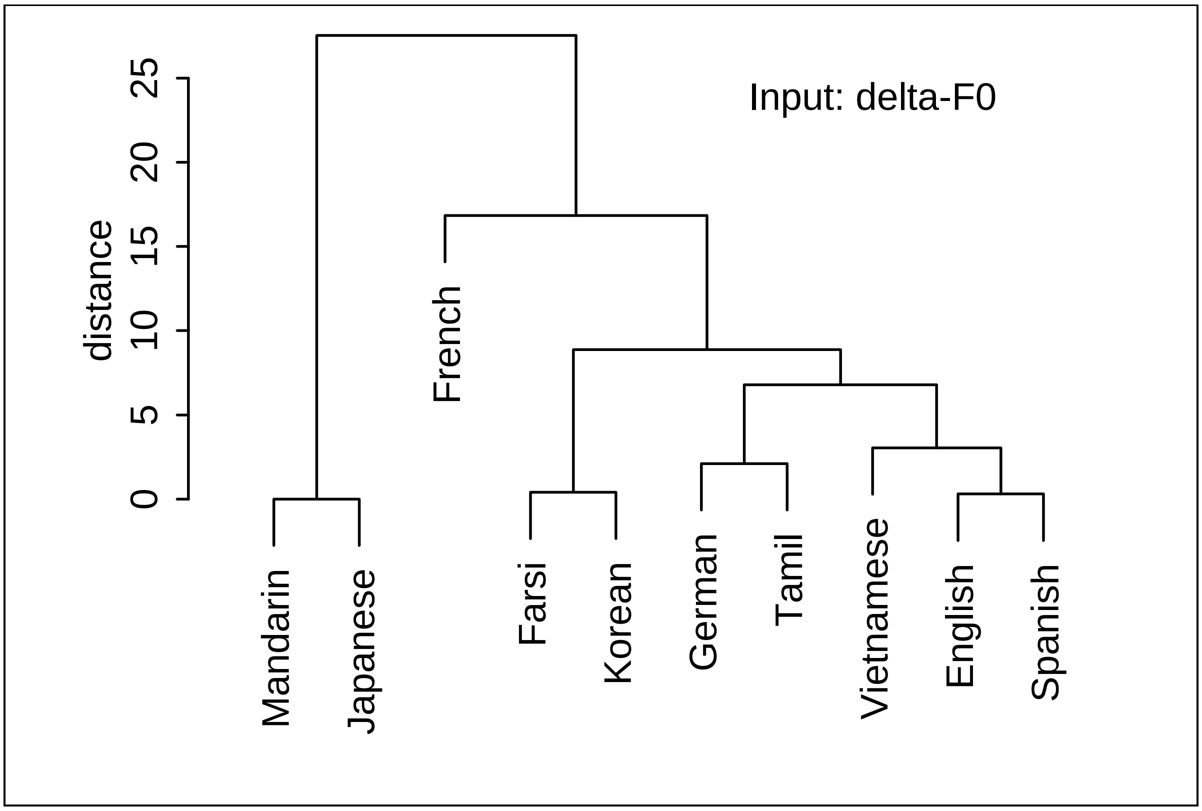

Figure 4:

4.2 Comparisons based on ΔF_0

The same training scheme was now applied using only ΔF_0 as an input. Table II gives the mean percent correct (s.d.). Overall, performance lies in the range 47.4% (Japanese–Mandarin) to 73.2% correct (French–Japanese). Figure 4 shows the resulting clusters.

Mandarin and Japanese appear to be indistinguishable when networks are only given ΔF_0 as input (replicating our previous finding, Cummins et. al, Submitted), but are readily distinguished from many of the remaining languages. Surprisingly, Vietnamese, which has lexical tone, is not identified as distinct from the Indo-European languages. French, on the other hand, is sufficiently different from the bulk of the languages to resist incorporation into any cluster at an early stage.

5 Comparison based on both ΔEnv and ΔF_0

Finally, 10 networks were trained on each pairwise comparison, providing both ΔF_0 and ΔEnv as input. Table III provides the results for each comparison, while Figure 5 shows the result of the clustering procedure.

	Ge	Sp	Ma	Ja	Fa	Ko	Ta	Vi	Fr
En	51.2 (1.1)	50.1 (1.6)	**62.8 (2.6)**	**61.0 (2.5)**	51.7 (2.4)	51.8 (3.2)	58.2 (3.4)	50.8 (4.1)	**61.2 (4.9)**
Ge	–	52.0 (1.0)	**63.2 (1.4)**	**66.6 (4.1)**	58.7 (1.8)	56.4 (2.5)	51.1 (0.9)	52.2 (3.5)	59.1 (1.8)
Sp		–	**68.6 (2.2)**	**68.3 (0.9)**	54.4 (1.5)	52.1 (3.1)	50.5 (1.1)	52.5 (2.2)	53.2 (1.6)
Ma			–	47.7 (3.3)	59.5 (1.1)	57.8 (3.5)	**61.6 (1.8)**	**65.7 (1.1)**	**73.0 (1.7)**
Ja				–	**64.8 (2.7)**	**61.7 (0.8)**	**65.1 (0.9)**	**71.1 (1.2)**	**73.2 (1.3)**
Fa					–	50.2 (1.3)	51.2 (2.2)	57.5 (1.5)	**65.6 (1.6)**
Ko						–	54.1 (1.6)	**61.1 (2.5)**	**61.9 (1.1)**
Ta							–	58.2 (3.5)	59.1 (1.7)
Vi								–	55.1 (1.8)

Table 2: Average percent correct (and s.d.) for pairwise comparisons based only on ΔF_0.

	Ge	Sp	Ma	Ja	Fa	Ko	Ta	Vi	Fr
En	55.6 (3.1)	45.8 (3.4)	58.7 (1.2)	**63.2 (4.3)**	**60.3 (3.6)**	**63.1 (3.7)**	**62.0 (2.6)**	58.0 (3.8)	59.0 (2.6)
Ge	–	50.5 (1.1)	**67.6 (1.7)**	**68.2 (2.6)**	59.4 (2.0)	59.3 (1.3)	59.5 (3.0)	57.1 (4.4)	**61.0 (1.4)**
Sp		–	**61.0 (2.9)**	**74.5 (1.9)**	53.2 (2.8)	55.1 (1.8)	47.4 (1.6)	55.6 (2.6)	52.0 (2.1)
Ma			–	52.5 (3.3)	**61.0 (1.5)**	**62.2 (2.3)**	**63.8 (2.9)**	**68.1 (1.9)**	**68.6 (5.8)**
Ja				–	**68.3 (2.2)**	**62.1 (2.1)**	**65.2 (2.6)**	**74.3 (2.4)**	**78.2 (2.4)**
Fa					–	58.1 (1.8)	**63.3 (1.9)**	**63.3 (2.4)**	**64.8 (1.6)**
Ko						–	55.5 (1.9)	**63.0 (3.0)**	**61.5 (2.0)**
Ta							–	57.9 (1.4)	56.6 (1.9)
Vi								–	59.6 (2.5)

Table 3: Average percent correct (and s.d.) for pairwise comparisons based ΔEnv and ΔF_0 together.

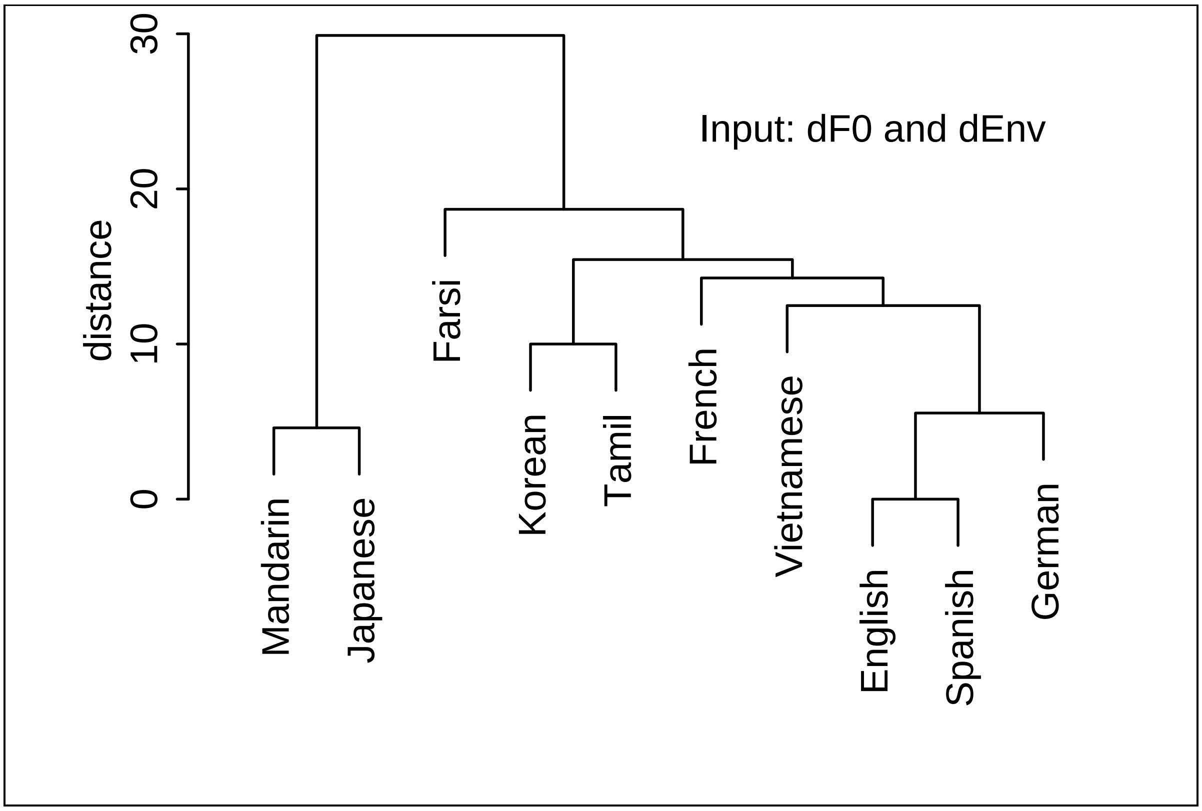

Figure 5: Clustering of languages based on both ΔEnv and ΔF_0.

Overall performance lies in the range 48.8% (English–Spanish) to 78.2% correct (Japanese–French).

Again we find Mandarin and Japanese clearly distinct from the remaining languages. Although mean discrimination at better than 70% was achieved using only ΔEnv, the two input task does no better than chance (52.5% correct, s.d. 3.3). This may be attributable to the increased difficulty of learning to ignore one, non-discriminatory input (ΔF_0), while simultaneously learning to exploit the other (ΔEnv). Our initial experience in deciding on architectural parameters, suggest that is unlikely that simply training a larger network would have produced any significant advantage here.

English, Spanish and German cluster together, but French remains distinct from this group, presumably by virtue of the marked intonational difference which was found in the preceding condition.

6 Discussion

In deliberating the failure of a simple rhythmic taxonomy to account for observed durational patterns in several languages, several authors have suggested that languages may differ along one or more continua, rather than forming discrete groups (Lehiste, 1977; Dauer, 1983; Nespor, 1990). A similar argument may be made for intonational types. Testing this hypothesis is difficult without further delineation of the supposed dimensions of variation. Dauer (1983, 1987) suggests that allowable syllable structures, degree of vowel reduction and phonetic manifestations of stress may all be variables contributing to perceived rhythmic type.

In the present approach, we manage to finesse the problem by assigning to the network the task of extracting relevant features from the input. By using several networks, each of which is started with a random weight initialization, we explore the space of possible discriminating functions for each pairwise comparison. Aggregate performance provides a distance metric between two languages, and similarities among languages and groups of languages are determined from the resulting distance matrix.

The approach has several shortcomings. In ignorance of the input features which actually distinguish a given pair of languages, we have no way of testing whether the network is doing as good a job as is possible. We must rather be content with observing that *some* features are being discovered which allow better than chance performance in many cases. Thus good discrimination is evidence for the existence of distinct features in the input, but failure to discriminate does not allow us to deduce the absence of discriminating features. Nevertheless, network performance is sufficiently reliable that we can identify both expected and some unexpected patterns in the resulting matrix. Expectations must, of course, be guided by the fact that the speech used is unscripted telephone speech from wide dialectal ranges, and not constrained laboratory speech.

We also deliberately restrict the information available to the network, by providing only ΔF_0 and/or ΔEnv as input. This prevents the network from assessing similarity on many possible bases, such as syllable complexity or vowel reduction, but allows direct assessment of the utility of the chosen variables in discrimination.

The clearest match between expectation and observation is obtained with ΔEnv as network input. Here, the distinction, often claimed but rarely supported, between syllable-timed and stress-timed languages is apparent in the clusters obtained. The failure to separate English and German from Japanese is unexpected, given that mora-timing is perhaps the only "rhythmic type" for which durational measurements actually offer strong support (Port et al., 1987). However, the regularity observed there was in fact non-local, in that mean mora duration remained constant, while the duration of individual morae varied substantially. It is possible, therefore, that the amplitude envelope modulation is, in fact, quite similar to that of a stress-timed language.

The finding of Ramus and Mehler (1999), however, that French natives could discriminate at better than chance levels between resynthesized stimuli based on English and Japanese, where the stimuli had all consonants replaced by /s/, all vowels by /a/ and a flat intonational contour, suggests that there is, indeed, some discriminating information in the amplitude envelope which the present networks are not sensitive to.

The present results strongly suggest that a simple continuum between stress-timed and syllable-timed languages will not suffice. Both Farsi and Vietnamese are quite distinct from the remain languages and from one another (70.9% correct). Unfortunately, we know of no phonetic studies of rhythm in these two languages which might suggest the manner in which they differ.

The observed utility of the first-differenced envelope of the band-limited signal stands in contrast to efforts to develop rhythm-based features of use in automatic language identification systems. Thymé-Gobbel and Hutchins (1996) concluded that "using amplitude features is a very poor LID strategy in that differences from speaker to speaker could cause discrimination results worse than chance". We, on the other hand, find ΔEnv to be a useful discriminant in many cases, and the only successful one in some (e.g. Mandarin-Japanese).

The overall discrimination performance obtained here is not as high as that of Thymé-

Gobbel and Hutchins (1996), who discriminated among a subset of the present languages using combinations of 224 predefined features. This suggests that there remain observable features of the data which are either no longer available to the network after input preprocessing, or to which the network is not sensitive. The latter may particularly be the case when the features in question are highly non-local, and dependent on the statistics of a large amount of speech. Although LSTM is capable of relating information separated by large temporal intervals, it is presumably easier to learn information which is local in time, such as the presence of a salient onset.

In the clusters obtained using two inputs, the separation of Mandarin and Japanese from the remaining languages, and the proximity of English, German and Spanish suggest that the combination of multiple inputs is likely to produce an empirical measure of "relatedness" among languages. Linguistic similarity is based, of course, on many more variables than the two applied here. There is nothing in principle to prevent the inclusion of further discriminating variables, either in raw, pre-linguistic form, as, e.g., linear predictive or cepstral coefficients, or as a series of temporally aligned symbols such as phonemes. This would allow an overall measure of phonetic similarity, but would lead significantly beyond the scope of the present work.

Future work will seek to optimize input representations, architecture and training parameters to obtain optimal performance in specific discrimination tasks. Optimization of all parameters was not possible in the present broad-band study, for obvious reasons of computational overhead. Networks which are found to discriminate optimally may reveal those features of the input which are actually useful for discriminating among languages. Thus, the present broadband approach can now be complemented by a detailed study of the informational basis for individual discrimination tasks.

7 Acknowledgements

Thanks to Cambridge University Engineering Department for providing the pitch_tracker program. This work was funded by SNF grant 21-49144.96 (Long Short-Term Memory).

References

Abercrombie, D. (1967). *Elements of general phonetics.* Aldine Pub. Co., Chicago, IL.

Balasubramanian, T. (1980). Timing in Tamil. *Journal of Phonetics*, 8:449–467.

Classé, A. (1939). *The Rhythm of English Prose.* Basil Blackwell, Oxford, England.

Cummins, F. (1997). *Rhythmic Coordination in English Speech: An Experimental Study.* PhD thesis, Indiana University, Bloomington, IN. Also Technical Report 198, Indiana University Cognitive Science Program.

Cummins, F., Gers, F., and Schmidhuber, J. (Submitted). Automatic discrimination among languages based on prosody alone. *Speech Communication.*

Cummins, F. and Port, R. F. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics*, 26(2):145–171.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, 11:51–62.

Dauer, R. M. (1987). Phonetic and phonological components of language rhythm. In *Proceedings of the 11th International Congress of Phonetic Sciences*, pages 113–121, Tallinn, Estonia. Academy of Sciences of the Estonian SSR.

Hazen, T. J. and Zue, V. W. (1994). Automatic language identification using a segment-based approach. In *Proceedings of the 3rd European Conference on Speech Communication and Technology*, pages 1307–1310.

Hochreiter, J. (1991). Untersuchungen to dynamischen neuronalen Netzen. Diploma thesis, Technische Universität München.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Itahashi, S., Zhou, J. X., and Tanaka, K. (1994). Spoken language discrimination using speech fundamental frequency. In *Proceedings of the 1994 International Conference on Spoken Language Processing*, volume 4, pages 1899–1902, Yokohama, Japan.

Kiparsky, P. and Youmans, G., editors (1989). *Rhythm and Meter*, volume 1 of *Phonetics and Phonology*. Academic Press, San Diego.

Lee, H.-B., Jin, N.-T., Seong, C.-J., Jung, I.-J., and Lee, S.-M. (1994). An experimental phonetic study of speech rhythm in Standard Korean. In *Proceedings of the 1994 International Conferences on Spoken Language Processing*, volume 3, pages 1091–1094, Yokohama, Japan.

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, 5:253–263.

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

Major, R. C. (1981). Stress-timing in Brazilian Portuguese. *Journal of Phonetics*, 9:343–351.

Mozer, M. C. and Smolensky, P. (1989). Skeletonization: A technique for trimming the fat from a network via relevance assessment. In Touretzky, D. S., editor, *Advances in Neural Information Processing Systems 1*, pages 107–115. San Mateo, CA: Morgan Kaufmann.

Muthusamy, Y. K., Cole, R. A., and Oshika, B. T. (1992). The OGI multi-language telephone speech corpus. In *Proceedings of the International Conference on Spoken Language Processing 1992*, Banff, Canada.

Nazzi, T., Bertoncini, J., and Mehler, J. (1998). Language discrimination by newborns: towards an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, 24:756–766.

Nespor, M. (1990). On the rhythm parameter in phonology. In Roca, I. M., editor, *Logical Issues in Language Acquisition*, pages 157–175. Foris, Dordrecht.

Pike, K. L. (1945). *The Intonation of American English.* University of Michigan Press, Ann Arbor, MI.

Pointon, G. E. (1980). Is Spanish really syllable-timed? *Journal of Phonetics*, 8:293–304.

Port, R. F., Dalby, J., and O'Dell, M. (1987). Evidence for mora timing in Japanese. *Journal of the Acoustical Society of America*, 81(5):1574–1585.

Ramus, F. and Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America*, 105(1):512–521.

Savithri, S. R. (1995). On speech rhythms in Kannada. *Journal of Phonetics*, 23:273–289.

Scott, S. K. (1993). *P-centers in Speech: An Acoustic Analysis.* PhD thesis, University College London.

Statistical Sciences (1995). *S-PLUS Guide to Statistical and Mathematical Analysis,Version 3.3.* StatSci, a division of MathSoft, Inc., Seattle.

Tajima, K. (1998). *Speech Rhythm in English and Japanese: Experiments in Speech Cycling.* PhD thesis, Indiana University, Bloomington, IN.

Thymé-Gobbel, A. and Hutchins, S. E. (1996). On using prosodic cues in automatic language identification. In *Proceedings of the 1996 International Conference on Spoken Language Processing*, volume 3, pages 1768–1771, Philadelphia, PA.

Wenk, B. J. and Wioland, F. (1982). Is French really syllable timed? *Journal of Phonetics*, 10:193–216.

Zissman, M. A. (1995). Language identification using phoneme recognition and phonotactic language modeling. In *Proceedings of the 1996 International Conference on Acoustics, Speech, and Signal Processing*, pages 3503–3506.