3
Prosody, Models, and Spontaneous Speech

Hiroya Fujisaki

ABSTRACT This paper presents a definition of prosody as the organization of linguistic units within an utterance and a coherent group of utterances, having manifestations both in segmental and suprasegmental features of speech, serving at the same time as a medium for conveying para- and nonlinguistic information. It then discusses the process of spontaneous speech production, emphasizing the role of quantitative generative models in both speech synthesis and speech recognition, examples are illustrated in Japanese. Finally, it discusses the continuum of spontaneity in speech, and briefly touches on the characteristics of speech that become dominant with increased spontaneity.

3.1 What is Prosody? Its Nature and Function

Although there are many views on prosody (e.g., Lehiste [Leh70]), I do not intend to elaborate on their differences here. These views may be broadly classified, after Ladd and Cutler [LC83], into the following two categories:

1. “Concrete”—defining prosody more or less in physical terms, as those phenomena that involve the acoustic parameters of pitch, duration, and intensity.
2. “Abstract”—defining prosody more from the point of view of its place in linguistic structure than its phonetic nature, as phenomena that involve phonological organization at levels above the segment.

According to Ladd and Cutler, these two approaches can be characterized—or caricatured—by their methodological preferences for either “taking measurements” or “building models.”

I consider that the study of human communicative behavior through language belongs to the empirical sciences, where one needs to obtain, first and foremost, clear and objective knowledge of the phenomena through

---

1Here I would use “theories” instead of “models,” since the word “model” is being used in a different meaning in this paper.
making measurements. Such knowledge can be more powerful if it is not confined to that obtained only from direct observations but can be generalized in the form of theories to cover instances that are not yet encountered or observed. By induction, i.e., going from measurements to theories, we can obtain the underlying, general principles. We can use intuition to build theories, i.e., to posit general, underlying principles first, but their validity has to be tested by confirming that their predictions agree with the measurements. The process of obtaining predictions for concrete, individual facts is through deduction. Thus the two processes—induction and deduction—are not dichotomous; both can be profitably used in the scientific quest for precise and generalizable knowledge.

Prosody has both measurable manifestations and underlying principles. It is only realized when a message is produced as a coherent string of sounds as speech. I would therefore like to present a third definition:

3. Prosody is the systematic organization of various linguistic units into an utterance or a coherent group of utterances in the process of speech production. Its realization involves both segmental and suprasegmental features of speech, and serves to convey not only linguistic information, but also paralinguistic and non-linguistic information.

Here I define linguistic information as the symbolic information that is represented by a set of discrete symbols and rules for their combination. It can be represented either explicitly by the written language, or can be easily and uniquely inferred from context. Linguistic information thus defined is discrete and categorical. For example, the information concerning the accent type of a Japanese word is discrete in the sense that it specifies one of a finite number of possible accent types.

On the other hand, paralinguistic information is defined as the information that is not inferable from the written counterpart but is deliberately added by the speaker to modify or supplement the linguistic information. A written sentence can be uttered in various ways to express different intentions, attitudes, and speaking styles which are under the conscious control of the speaker. Paralinguistic information can be both discrete and continuous. For example, the information regarding whether a speaker's intention is an assertion or a question is discrete, but it can also be continuous in the sense that a speaker can express the degree within each category.

---

2I am aware that these definitions of linguistic and paralinguistic information are unconventional (e.g., Crystal [Cry69], Laver [Lav94]). However, they are introduced here to deal more systematically with various functions of prosody and prosodic features than conventional definitions do.

3It is possible that a speaker's intention can be expressed either as linguistic information or as paralinguistic information, or both. For instance, interrogation can be regarded as linguistic (in the sense defined above) if a speaker uses