Communication and Prosody: Functional Aspects of Prosody

Julia Hirschberg
AT&T Labs — Research
A257 — Shannon Laboratory
Florham Park, NJ 07932-0971
julia@research.att.com
http://www.research.att.com/ julia/index.html

Number of pages: 41.

Keywords: prosody, spoken dialogue systems

Abstract

interest in the contribution prosodic information makes to human communication has led to increasing expectations that such information could be of use in text-to-speech and speech understanding systems, and in application of these technologies to spoken dialogue systems. To date, research results far exceed their technology applications. This paper suggests some areas in which progress has been made, and some in which more might be made, with particular emphasis upon text-to-speech synthesis and spoken dialogue systems.

1 INTRODUCTION

In the past decade, there has been a growing appreciation of the important role of prosody in human-human as well as in and in human-machine communication. Linguists, computational linguists and speech engineers have increasingly looked to intonation as an important component in language processing: Syntacticians and semanticists often appeal to prosody to disambiguate structural or scope ambiguities. Students of pragmatics and discourse look for prosodic cues to the conveyance of direct vs. indirect speech acts or hierarchical discourse structure. Text-to-speech systems compete to improve prosodic assignment and realization to produce more ”natural” utterances.

These years have seen some pronounced trends also in developing new schemes for prosodic description, such as the ToBI system described below, to allow researchers to compare their findings more easily, within and across languages, and to facilitate the construction of very large labeled speech corpora, especially for learning associations between prosodic features and other aspects of the text. In fact, corpus-based prosodic research has become quite important, especially for speech technologists. Also perhaps due to influence from the applications areas for prosodic research, the evaluation of claims about prosodic phenomena and algorithms developed to assign or generate prosodic features, for example, have received considerable attention. There are new questions, for example, on how to evaluate prosodic assignment for text-to-speech applications: should systems attempt to model the observed prosodic variations of a particular speaker, or should they simply be expected to produce prosody that might plausibly be generated by some native speaker of the language? Spontaneous speech phenomena have also attracted a considerable amount of attention among researchers on prosody, with attempts made to characterize filled pauses and speech disfluencies prosodically, primarily for purposes of improving speech recognition.

For many years the application focus of prosodic research has been limited primarily to text-to-speech — learning how to assign intonational variation from an analysis of input text, as well as how to realize that assignment in the output speech. Today, we see increasing interest in prosodic research in other technologies, like automatic speech recognition. Speech researchers are increasingly seeking to make use of prosodic information to reduce perplexity in search in automatic speech recognition, to disambiguate lexical choice, and to enhance general language understanding.

Spoken dialogue systems, drawing upon both speech generation and speech understanding technologies, also present their own particular needs and opportunities for prosodic investigation. In this paper, we will identify some basic areas of prosodic variation and some of their functions, review current directions in research on prosody, focussing on general issues surrounding the role of prosody in human-human communication, note some current issues and future directions in the study of prosody for speech generation, for speech understanding, and for the special application of these to spoken dialogue systems.

2 FUNCTIONS OF PROSODIC VARIATION

While much has been learned about intonational meaning in recent years, results have often been slow to find their way into speech applications, such as text-to-speech systems, concept-to-speech systems, and spoken dialogue systems. This section provides a tutorial overview of some of the major areas of research on prosody, with some pointers to work in each area, and with specific reference to their role in text-to-speech systems.[1] We will also note those aspects of prosodic research which seem most promising for incorporation into speech technologies.

Before proceeding to survey the intonational literature particularly relevant to spoken dialogue systems, however, two caveats should be made. First, there are many ways to say very similar things, whether through prosodic means or other linguistic behavior. Focus, for example, can be conveyed through intonational prominence or through intonational phrasing variation, or through variation in word order or the use of particular syntactic constructions. There is no single method a given speaker employs to convey a particular kind of meaning, and there is certainly no single method all speakers use to convey such meaning. So, research on prosody, as on many linguistic phenomena which rely upon context for their interpretation, is more a matter of finding likelihoods — not simple mappings from syntax or semantics or even from an underlying meaning representation to a clear set of prosodic features, for any sentence. Corpus-based research though, has considerable risks, since the "gold standard" of prosodic performance for an individual utterance is quite elusive. In text-to-speech systems, it is probably wisest to model the variability of a single speaker, but obtaining a large enough labeled corpus from one speaker to capture the full range of prosodic, syntactic, and semantic meanings possible for unrestricted text is fairly daunting. Second, the same prosodic feature can be used to communicate many different meanings. An expansion of a speaker's pitch range, for example, can convey different interpretations of a single intonational contour, a change in the speaker's degree of involvement with a subject, a shift in topic, or a return from a parenthetical remark. Since the way these different functions interact has not been systematically studied, it is particularly difficult in text-to-speech systems, or in speech understanding systems, to determine how to compose or to decompose a prosodic feature properly. With both these cautions in mind, we can still find some regularities in prosodic behavior that we can build upon.

2.1 The ToBI Intonational Model

To discuss prosodic variation from either a scientific or an applications point of view, it is first necessary to choose some framework of intonational description to talk about prosodic phenomena within. The intonational model which will be used to describe prosodic phenomena below is the ToBI model for standard American English (Pitrelli, Beckman, and Hirschberg, 1994; Silverman et al., 1992).[2] Claims made should be evaluated with respect to standard American English only.

The ToBI system consists of annotations at four, time-linked levels of analysis: an ORTHOGRAPHIC TIER of time-aligned words; a BREAK INDEX TIER indicating degrees of junction between words, from *0* 'no word boundary' to *4* 'full intonational phrase boundary' (Price et al., 1990); a TONAL TIER, where PITCH ACCENTS, PHRASE ACCENTS and BOUNDARY TONES describing targets in the FUNDAMENTAL FREQUENCY (f0) define intonational phrases, following Pierrehumbert's (Pierrehumbert, 1980) scheme for describing American English with modifications.

Break indices define two levels of phrasing, minor or intermediate (level 3) and major or intonational (level 4), with an associated tonal tier that describes the phrase accents and boundary tones for each level. Level 4 phrases consist of one or more level 3 phrases, plus a high or low boundary tone (**H%** or **L%**) at the right edge of the phrase. Level 3 phrases consist of one or more pitch accents, aligned with the stressed syllable of lexical items, plus a phrase accent, which also may be high (**H-**) or low (**L-**). A standard declarative contour, for example, ends in a low phrase accent and low boundary tone, and is represented by **L-L%**; a standard yes-no-question contour ends in **H-H%**. Five types of pitch accent occur in the ToBI for American English: two simple accents (**H*** and **L***, and three complex ones, **L*+H**, **L+H***, and **H+H***. As in Pierrehumbert's system, the asterisk indicates which tone is aligned with the stressed syllable of the word bearing a complex accent.

2.2 Contour Variation

There is a rich linguistic tradition characteristizing variation in overall pitch contour in many different ways: syntactic mood, speaker attitude, speaker beliefs (Bolinger, 1986; Bolinger, 1989; Ladd, 1980; Ladd, 1996). Some inherent meaning has often been sought in particular contours — often modulated by context (Liberman and Sag, 1974; Sag and Liberman¡, 1975;

Ladd, 1977; Ladd, 1978; Bing, 1979; Ladd, 1980; Bouton, 1982; Ward and Hirschberg, 1985; Grabe et al., 1997; Gussenhoven and Rietveld, 1997). And more general attempts have been made to identify compositional meanings for contours within various systems of intonational analysis (Gussenhoven, 1983; Pierrehumbert and Hirschberg, 1990). Linguists often seek to define 'standard' contours for declaratives, wh-questions, yes-no-questions.[3] Phrases ending in **L-H%** (often called "continuation rise"), convey the impressions of their being "more to come" (Bolinger, 1989); **L*+H** accents combined with continuation rise (the *Rise-Fall-Rise contour*) produces the effect of uncertainly or incredulity, depending upon pitch range, rate, and amplitude (Hirschberg and Ward, 1992); the *Plateau contour* consists of **H*** accents with H-L% and conveys a somewhat bored, recitation effect — "You already know this and I'm just reminding you of it"; **H*** accents with H-H% (the *High-Rise Question contour*) convey a subtle form of appropriateness, rather than a content question; and many more meaningful contours have been identified and investigated or speculated upon.

Varying contour appropriately in text-to-speech systems has traditionally been confined to attempts to assign contour appropriately for declaratives and questions, whose identity is inferred from sentence-final punctuation and the presence of *wh*-words, and to employ continuation rise at non-final punctuation. While it seems more likely that concept-to-speech systems would perform well at contour variation, in fact, little work in this area has been done, perhaps because the contours that have been well studied, such as Rise-Fall-Rise, express meanings that such systems may not care to produce.

2.3 Variation in Location and Type of Pitch Accents

Pitch accents make items intonationally prominent, and this prominence can be achieved via different tone targets, as well as differences in f0 height, to convey different messages (Terken, 1997; Campbell and Beckman, 1997). So, items may be accented or not (DEACCENTED (Ladd, 1979b)), and, if accented, may bear different tones, or different degrees of prominence, with respect to other accents (Terken, 1997). The perceptually most prominent accent in a prosodic phrase is generally known as its NUCLEAR STRESS. Constraints on nuclear (sometimes termed sentence) stress are discussed in (Cutler and Foss, 1977; Erteschik-Shir and Lappin, 1983; Schmerling, 1976; Schmerling, 1974; Bardovi-Harlig, 1983b). Despite Bolinger's seminal article on the unpredictability of accent (Bolinger, 1972), attempts to do so from related features of the uttered text continue (Altenberg, 1987; Hirschberg, 1993), especially for accent assignment in text-to-speech.

Certain lexical categories appear to have different propensities for accentuation than others, a fact frequently made use of in text-to-speech systems. So function words tend to be deaccented and content words, accented. Within these broad classes, however, differences abound. Particles and verbal prepositions, for example, such as *up* in *back up the disk*, tend to be accented far more often than prepositions in similar positions. However, prepositions too may be accented, to convey focus or contrast, as in *I didn't shoot AT him, I shot PAST him.* And there is some evidence that items interpreted as narrowly focussed or contrastive represent different accent types, and not simply differences in relative prominence (Krahmer and Swerts, 1998).

While pronouns tend to be deaccented, they can be accented to convey various ‘marked’ effects, as in the classic example due to Lakoff (Lakoff, 1971). When *he* and *him* are deaccented in *John called Bill a Republican and then he insulted him* the inferred resolution of referents is typically different than when both are accented. Similarly, the interpretation of the second clause in *John likes his colleagues and so does Sue* can be affected depending upon the accentuation of the pronoun in the first. With no accent on *his*, listeners are likely to understand that ‘John likes his colleagues and Sue also likes John’s colleagues’. Whereas, with *his* accented, listeners are likely to understand that ‘John likes his own colleagues and Sue likes her own colleagues’. Accent can also disambiguate potentially ambiguous words such as DISCOURSE MARKERS, or CUE PHRASES, words and phrases such as *now*, *well*, *in the first place*, which can either serve as explicit indicators of discourse structure or can have a *sentential* reading, often as adverbials. For example, if *now* is realized with a high pitch accent in *Now Bill is a vegetarian*, it is more likely to be interpreted in an adverbial sense; deaccented or with a **L*** accent, it is more likely to be interpreted in its discourse particular sense. For a spoken dialogue system, it is important to realize such phenomena so as to convey the intended meaning. For example, if *now* is to be interpreted as a temporal adverbial, it should be given an **H*** accent and should be part of the larger intonational phrase in (1).

(1) System: Now let me get you the train information.

If instead it should be interpreted as a discourse marker, it might be realized with a **L*** accent, or set apart from the remaining material as a separate intonational phrase.

A number of authors have examined the relationship between accent and various types of INFORMATION STATUS, including THEME/RHEME, TOPIC/COMMENT, and GIVEN/NEW status (Schmerling, 1975; Bardovi-Harlig, 1983a; Brown, 1983; Gundel, 1978; Lehman, 1977; Fuchs, 1980; Chafe, 1976; Nooteboom and Terken, 1982; Fuchs, 1984; Terken, 1984; Terken, 1985; Terken and Nooteboom, 1987; Fowler and Housum, 1987; Horne, 1991a; Horne, 1991b; Allerton and Cruttenden, 1979; Kruyt, 1985; Cahn, 1998; Terken and Hirschberg, 1994). It is a common generalization that speakers typically deaccent items that represent old, or GIVEN information in a discourse (Prince, 1981). However, the number of exceptions to this assumption, and the difficulty often of defining what ‘givenness’ is, have made this a fertile subject for research and experimentation. Whether or not a ‘given’ item participates in a complex nominal, the location of such an item in its prosodic phrase and whether preceding items in the phrase are ‘accentable’ due to their own information status, the grammatical function of an item when first and subsequently mentioned — all affect whether or not a ‘given’ item is deaccented or not.

Given/new status is modelled in text-to-speech systems at best by collecting stems of previously uttered items in a fixed window, or a paragraph or other orthographic unit, and counting those items as ‘given’, and hence, DEACCENTABLE. This simple procedure tends to deaccent too many items, based as it is on a simple notion of givenness, and without much sensitivity to the factors which interact with it. In concept-to-speech (Gawronska and House, 1998), of course, this feature is potentially under greater system control, although some algorithm for specifying what is treated as new must be specified. In spoken dialogue systems, however, this simple definition should prove a more reliable guide to speech production: i.e., items previously mentioned by system or user in the dialogue, should be considered 'given' in subsequent turns. Of course, it is not yet clear whether what is 'given' for a speaker, should also be treated as 'given' for his/her illocutionary partner — and thus, deaccentable. For example, while it seems plausible that *return* might be deaccentable (as 'given') in example (2), empirical results from the Edinburgh Map tasks dialogues suggest that such clearly 'given' items are rarely deaccented across speakers:(Bard, 1999)

(2) System: Do you want a return ticket?

User: No, thanks. I don't need a return.

Also, there are cases where repeated information is clearly not 'given' (Shimojima et al., 2001).

Changing the accent pattern of an utterance by accenting some words and failing to accenting others, can change the meaning of an utterance dramatically. For example, in the classic example *John only introduced Mary to Sue*, with the word *Mary* given nuclear stress, the utterance is likely to convey that Mary is the only person John introduced to Sue; but with *Sue* receiving nuclear stress, Sue is the only person John introduced Mary to. These differences are often called differences in FOCUS and may be tested by asking: What question is this utterance a felicitous answer to? The answer to this question is generally the focus of the original utterance. FOCUS-SENSITIVE OPERATORS, such as *only*, which interact with intonational prominence to produce variation in focussing effects, include other quantifiers (*all*, *most*, *some*), adverbs of quantification (*sometimes most often*), modals (*must*), emotive factives/attitude verbs (*It's odd that*), counterfactuals, and various other constructions. Work on the focal domains of accent and the representation and interpretation of intonational focus and presupposition includes (Lakoff, 1971; Schmerling, 1971; Jackendoff, 1972; Ball and Prince, 1977; Wilson and Sperber, 1979; Enkvist, 1979; Gussenhoven, 1983; Culicover and Rochemont, 1983; Rooth, 1985; Rochemont and Culicover, 1990; Rooth, 1991; Horne, 1985; Horne, 1987; Baart, 1987; Dirksen, 1992; Zacharski, 1992). Taking advantange of focal information in any speech technology has proven difficult; for text-to-speech, it requires independent access to information about what is to be focussed; for speech understanding systems, it is difficult, as noted above, to determine just why some item has been given particular intonational prominence. In concept-to-speech systems, however, which typically mark items to be focussed internally, it is simpler to utilize such information effectively in production (Horne and Filipsson, 1994; Williams, 1998). The question remains, of course, of just how to realize focus — through intonational means or otherwise.

While we have so far treated accent as a binary feature in this section, as noted in Section 2.1, there are differences in the type of pitch accent a lexical item is associated with, as well as the relative prominence of that accent within an intermediate phrase. Differences in accent type convey considerable differences in meaning in conjunction with differences in the discourse context and variation in other acoustic properties of the utterance. **H*** accents are the commonest in English and are found in standard declarative utterances, while **L*** accents characterize yes-no question contours.[4] **L+H*** accents appear to mark items as contrastive or narrowly focussed, e.g. 'I want to go to **L+H*** Boston, not **Baltimore**'. More 'scooped' accents are **L*+H**, which may be interpreted as conveying uncertainty or incredulity, depending on the pitch range and voice quality associated with their utterance, e.g. ‘I **L*+H** thought you said Baltimore’ vs. ‘But you **L*+H** said Baltimore.’ And **H+!H*** accents, realized as a fall onto the stressed syllable, are associated with some implied sense of familiarity with the mentioned item.

2.4 Phrasing Variation

Intuitively, prosodic phrases divide an utterance into meaningful ‘chunks’ of information (Bolinger, 1989). Appropriate ‘chunking’ has been found to be important to comprehension and perceived naturalness (Sanderman and Collier, 1997). Both level 3 (intermediate) and level 4 (intonational) phrases are identified by changes in f0, and frequently associated with other acoustic and prosodic cues, such as PHRASE-FINAL LENGTHENING, glottalization (‘creaky voice’) over the last syllable or syllables in the phrase, and some amount of pause. Not all perceived phrase boundaries exhibit all features, but in general, level 4 boundaries tend to exhibit more pronounced cues than level 3.

Variation in phrasing can change the meaning hearers assign to a sentence.[5] For example, the interpretation of a sentence like *Bill doesn't drink* {|} *because he's unhappy* is likely to change, depending upon whether it is uttered as one phrase (wide scope negation: Bill does indeed drink — but the cause of his drinking is not his unhappiness.) or two (narrow scope: Bill's unhappiness has lead him **not** to drink). There are many other constructions in which phrasing appears to exhibit syntactic correlations, and thus to serve a potentially disambiguating function.

Phrasing can distinguish among different readings of semantically ambiguous utterances such as the scope of negation ambiguity in (3):

(3) a. I don't travel by ship | because I'm too cheap. [I don't travel by ship]

b. I don't travel by ship because I'm too cheap. [I travel by ship, but not because it's cheaper]

And it can also disambiguate the attachment of ambiguous constituents, such as *pp*s (4) and relative clauses (5).

(4) a. Contact the ticket office | in Baltimore. [when you are in Baltimore contact the ticket office]

b. Contact the ticket office in Baltimore. [contact the ticket office which is in Baltimore]

(5) a. The next train which is going to Baltimore is the one you want. [you want the next train going to Baltimore]

b. The next train | which is going to Baltimore | is the one you want. [you want the next train]

And the scope of modifiers can also be disambiguated by variation in prosodic phrasing, as in (6):

(6) a. This fare is restricted to retired school teachers | and civil servants. [all civil servants can get this fare]

b. This fare is restricted to retired | school teachers and civil servants. [only retired civil servants can get this fare]

While intonational phrasing **can** serve all these functions, from both corpus-based studies (Altenberg, 1987; Ostendorf and Veilleux, 1994; Hirschberg and Prieto, 1996; Fujio, Sagisaka, and Higuchi, 1997), or laboratory experiments (Grosjean, Grosjean, and Lane, 1979; Wales and Toner, 1979; Gee and Grosjean, 1983; Price et al., 1990; Beach, 1991; Hirschberg and Avesani, 1997), evidence that it does so reliably is mixed. Speakers rarely recognize the potential ambiguity of the sentences they utter, and routinely violate most of the distinctions illustrated above. So, although much interest in defining a clear mapping between prosody and syntax has persisted through the years, both in linguistic and engineering circles (Downing, 1970; Bresnan, 1971; Selkirk, 1984; Cooper and Paccia-Cooper, 1980; Dirksen and Quene, 1993; Prevost and Steedman, 1994; Boula de Mareüil and d'Alessandro, 1998), it is important to temper our expectations of how intonational phrasing information can best be employed, in text-to-speech, concept-to-speech, speech understanding, and general dialogue systems applications.

First, as the examples above illustrate, when text itself is ambiguous, appropriate boundary location is especially difficult for text-to-speech systems, although possible for concept-to-speech (Klabbers, Krahmer, and Theune, 1998). For text-to-speech, due to the considerable variability in human performance, the goal of phrasing modules is likely to remain one of avoiding what are clearly errors (i.e. phrasing no human being would produce) and apparent disfluencies (i.e. hesitations which humans **do** produce but do not perhaps aim for in public speech). Syntactic or other cues to such errors might in fact be a more useful subject of study for text-to-speech research than fluent human performance. This seems likely to be the explanation for the success of intonational phrasing experiments in reducing perplexity in speech recognition tasks (Hess et al., 1997) — they disfavor very unlikely combinations of syntactic context and prosodic phenomena.

2.5 Varying Timing and Pitch Range

Variation in aspects of pitch range as well as rate can change the meaning of particular intonational contours, such as the rise-fall-rise contour (**L*+H L-H%**), as noted in Section 2.2. Range variation can also convey differences in degree of speaker 'involvement'; expanded pitch range seems to communicate a greater degree of involvement. Rate, duration of inter-phrase pause, loudness, and pitch range can also convey the topic structure of a text (Silverman, 1987; Avesani and Vayra, 1988; Grosz and Hirschberg, 1992; Ayers, 1992; Swerts, Collier, and Terken, 1994; Swerts, 1997; Brown, Currie, and Kenworthy, 1980; Lehiste, 1979; Avesani and Vayra, 1988; Passoneau and Litman, 1993; Hirschberg and Nakatani, 1996; Koiso, Shimojima, and Katagiri, 1998; van Donzel, 1999). In general, various researchers have found that phrases beginning new topics are begun in a wider pitch range, are preceded by a longer pause, are louder, and are slower, than other phrases; narrower range, longer subsequent pause, and faster rate characterize topic-final phrases. Subsequent variation in these features then tends to be associated with a topic shift.

While such results have been widely disseminated, it has proven difficult to take advantage of them — either in text-to-speech, due to difficulties of identifying topic structure from text, or in automatic speech recognition, where segmentation based upon lexical cues has so far been a more popular approach. And, although important correlations have been found between acoustic features and topic structure, it is hard to reduce these descriptive findings to a recipe for production or perception.

Range and rate can also distinguish phenomena such as parenthetical phrases from others: parentheticals are generally uttered in a compressed pitch range and with a faster speaking rate than other phrases. And initial phrases of direct quotations are uttered in an expanded pitch range. FINAL LOWERING, a compression of the pitch range during the last half second or so of an utterance, can also convey structural information to hearers, by signalling whether or not a speaker has completed his/her 'turn'. Pitch contour and range as well as timing have also been shown to correlate with turn-final vs. turn-keeping utterances — and distinguishing the former from discourse boundaries — as well as marking backchannels in dialogue (Geluykens and Swerts, 1994; Koiso et al., 1998; Caspers, 1998).

3 INTONATIONAL VARIATION IN SPOKEN DIALOGUE SYSTEMS

To date there has been little direct application of intonational research results to the development of spoken dialogue systems, except insofar as improving the prosody of text-to-speech and utilizing prosodic information in speech recognition improves dialogue systems' component parts. More certainly might be done in both these technology areas.

For text-to-speech systems, much might be done now to incorporate research findings into prosodic prediction procedures. Evidence that grammatical function is an important factor in determining whether 'given' items are accented or not could improve the tendency of systems with simpler assumptions about given/new to deaccent too much. The vast number of findings on the correlates of discourse structure might be made use of in text-to-speech as well as message-to-speech, to improve the production of paragraphs, for example, by varying range, pausal duration, and rate appropriately. Faster and better parsers exist than are currently used in text-to-speech systems, whose output might improve the assigment of prosodic phrase boundaries. And major architectural as well as scientific issues about how to integrate the prediction of accent with the prediction of phrasing still need to be addressed. For message-to-speech systems, larger questions of how prosodic specification is made together with decisions about lexical and syntactic realization remain relatively untouched. And despite a large amount of interest in and research on the prosodic correlates of emotional speech (Cahn, 1988; Murray and Arnott, 1993; Schröder, Auberge, and Cathiard, 1998; Whiteside, 1998; Koike, Suzuki, and Saito, 1998; Rank and Pirker, 1998; Amir and Ron, 1998) and of individual speaking style, attempts to develop different 'voices' or styles (Abe, 1997) in text-to-speech systems are still primarily of curiosity value rather than real choices for system developers.

For spoken dialogue systems, however, there are additional questions to be addressed:

Prosody could be used more effectively to convey information that currently is lexicalized, to decrease the length of system responses or improve naturalness. For example, system confirmations could rely more upon yes-no question intonation to eliminate redundant material, as asking 'Baltimore?' instead of 'Did you say you wanted to go to Baltimore?' Since confirmations can be quite annoying when the system has correctly understood user input, this greater terseness might address that problem. Certainty and uncertainty,(Gorin, 1995) questioning behavior, politeness, all could be effectively conveyed by the use of appropriate contours, given a text-to-speech system which can realize these well.

Prosodic correlates discovered for human-human turn-taking behavior could be employed more effectively to signal when the system is keeping the floor and when it wishes to relinquish it. Much confusion in spoken dialogue systems arises when users become confused about when they are expected to supply input, and when the system is still processing their prior utterance and not listening for a new one. While more explicit techniques than those humans employ to take and relinquish turns will probably be required to fully address this problem, current systems would be well advised **not** to inadvertently signal end-of-turn, as, by ending a mid-turn utterance like 'The next train to Boston leaves at 8:30 p.m.' with falling intonation in the turn illustrated in (7):

(7) System: The next train to Boston leaves at 8:30 p.m. There is a faster train at 8:45 though.

Users are more likely to wait for the additional information about the faster train, if the fact that there is 'more to come' is signalled by the use of continuation rise on *p.m.*.

A more general use of prosodic variation in spoken dialogue systems parallels the effort to find user-friendly voice talents and appropriate speaking styles for interactive systems that use stored-voice prompts rather than text-to-speech for system generation. Systems which do employ text-to-speech output would be well advised to set pitch range, rate, and other prosodic parameters in the same way that voice talent is currently chosen for stored-speech systems. No single voice or set of parameter settings in any text-to-speech system will necessarily be right for all applications and all target markets. The findings of (Silverman et al., 1993) that tuning a poorly-evaluated synthesizer's prosodic parameters to match those of humans actually performing a reverse telephone directory lookup turned the least effective system into the most effective is a clear sign of how important this neglected aspect of dialogue systems can be. In the longer term, dialogue systems might even aspire to adapting a system's prosodic parameters to more closely match those of users, to make the system seem more familiar, with respect to features such as pitch range or rate, with current technology for analysis and production.

Another important use of prosodic that is only beginning to be explored focusses upon the interpretation of user prosody in such systems. There is already considerable evidence suggesting that user corrections of system errors are often signalled by hyperarticulation — a slowing of rate, increase in loudness, and rise in overall pitch, as in (8):(Wade, Shriberg, and Price, 1992; Oviatt et al., 1996; Swerts and Ostendorf, 1997; Levow, 1998; Bell and Gustafson, 1999)

(8) User: I want to go to Baltimore on July 25th.
System: When do you want to go to Boston?
User: I said BAL-TI-MORE!

And recently we have discovered that prosodic features such as pitch, amplitude, and timing can actually signal utterances which will be misrecognized (Hirschberg, Litman, and Swerts, 1999).

Finally, again from the understanding perspective, there has been considerable effort to identify speech/dialogue acts using prosodic as well as lexical cues (Tamoto and Kawabata, 1998; Shriberg et al., 1998; Taylor et al., 1998; Warnke et al., 1997). Currently these efforts are focussed on improving automatic speech recognition performance, but in future it will be useful to investigate how systems' dialogue strategies (e.g. confirmation strategies, type of initiative) might be modified in light of such information (Kompe et al., 1994). Also, there is some evidence that prosody can be used to identify more 'salient' or interpretively useful portions of user input; for example, accented information may be more reliable information for overall utterance interpretation than unaccented information (Nöth et al., 2000). And the processing of short turns like *oh*, *okay*, and *well*, which can represent multiple speech acts, as users intend them must necessarily draw upon intonational information.

4 Conclusion

The importance of generating and recognizing prosodic information in spoken dialogue systems is only beginning to be explored. It is not sufficient for such systems simply to rely upon continuing advances in the use of prosody in their various component technologies, text-to-speech systems and speech recognizers. Particular problems both dialogue systems and their users face, for example, in understanding the state of their conversational partner — "Is this system working on my problem or has it died?", "Is this user finished with their input or should I wait for more?" — are most naturally solved by intonational means. However, unless the component technologies provide more sophisticated capabilities in both the generation and the recognition of prosodic variation, these needs cannot be addressed.

References

Abe, Masanobu. 1997. Speaking styles: Statistical analysis and synthesis by a text-to-speech system. In Jan P. H. van Santen, Richard Sproat, Joseph P. Olive, and Julia Hirschberg, editors, *Progress in Speech Synthesis.* Springer, pages 495–510.

Allerton, D. and A. Cruttenden. 1979. Three reasons for accenting a definite subject. *Journal of Linguistics*, 15(1):49–53.

Altenberg, Bengt. 1987. *Prosodic Patterns in Spoken English: Studies in the Correlation between Prosody and Grammar for Text-to-Speech Conversion*, volume 76 of *Lund Studies in English.* Lund University Press, Lund.

Amir, N. and S. Ron. 1998. Towards an automatic classification of emotions in speech. In *Proceedings of ICSLP-98*, Sydney. International Conference on Spoken Language Processing.

Avesani, Cinzia and Mario Vayra. 1988. Discorso, segmenti di discorso e un' ipotesi sull' intonazione. In *Corso di stampa negli Atti del Convegno Internazionale "Sull'Interpunzione"*, pages 8–53, Vallecchi, Firenze.

Ayers, Gayle M. 1992. Discourse functions of pitch range in spontaneous and read speech. Presented at the Linguistic Society of America Annual Meeting.

Baart, J. L. G. 1987. *Focus, Syntax and Accent Placement.* Ph.D. thesis, University of Leyden, Leyden.

Ball, C. N. and E. F. Prince. 1977. A note on stress and presupposition. *Linguistic Inquiry*, 8(3):585.

Bard, Ellen. 1999. The dissociation of deaccenting, givenness, and syntactic role in spontaneous speech. In *Proceedings of ICPhS99*, San Francisco, August. International Congress of Phonetic Sciences.

Bardovi-Harlig, K. 1983a. Pronouns: When 'given' and 'new' coincide. In *Papers from the 18th Regional Meeting*. Chicago Linguistic Society.

Bardovi-Harlig, Kathleen. 1983b. *A Functional Approach to English Sentence Stress*. Ph.D. thesis, University of Chicago, Chicago IL.

Beach, Cheryl. 1991. The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations. *Journal of Memory and Language*, 30:644–663.

Bell, Linda and Joakim Gustafson. 1999. Repetition and its phonetic realizations: Investigating a Swedish database of spontaneous computer-directed speech. In *Proceedings of ICPhS-99*, San Francisco. International Congress of Phonetic Sciences.

Benoit, Christian and Gerard Bailly, editors. 1989. *Proceedings of the European Speech Communication Association Workshop on Speech Synthesis*, Autrans, September. European Speech Communication Association.

Bing, Janet. 1979. *Aspects of English Prosody*. Ph.D. thesis, University of Massachusetts at Amherst, Amherst MA. Distributed by the Indiana University Linguistics Club.

Bolinger, Dwight. 1972. Accent is predictable (if you're a mindreader). *Language*, 48:633–644.

Bolinger, Dwight. 1986. *Intonation and Its Parts: Melody in Spoken English.* Stanford University Press, Palo Alto CA.

Bolinger, Dwight. 1989. *Intonation and Its Uses: Melody in Grammar and Discourse.* Edward Arnold, London.

Botinis, A., G. Kouroupetroglou, and G. Carayiannis, editors. 1997. *Intonation: Theory, Models and Applications*, Athens.

Boula de Mareüil, Philippe and Christophe d'Alessandro. 1998. Text chunking for prosodic phrasing in french. In *The Third ESCA/COCOSDA Workshop on Speech Synthesis*, pages 127–131, Jenolan Caves Mountain House, Blue Mountains, Australia, November.

Bouton, Lawrence F. 1982. Stem polarity and tag intonation in the derivation of the imperative tag. In Robinson Schneider, Kevin Tuite, and Robert Chametzky, editors, *Papers from the Parasession on Nondeclaratives*, pages 23–42, Chicago. Chicago Linguistic Society.

Bresnan, Joan. 1971. Sentence stress and syntactic transformations. *Language*, 47:257–281.

Brown, G. 1983. Prosodic structure and the given/new distinction. In D. R. Ladd and A. Cutler, editors, *Prosody: Models and Measurements.* Springer Verlag, Berlin, pages 67–78.

Brown, G., K. Currie, and J. Kenworthy. 1980. *Questions of Intonation.* University Park Press, Baltimore.

Cahn, J. 1988. From sad to glad: Emotional computer voices. In *Proceedings of Speech Tech '88*, pages 35–36.

Cahn, Janet. 1998. Generating pitch accent distributions that show individual and stylistic differences. In *The Third ESCA/COCOSDA Workshop on Speech Synthesis*, pages 121–126, Jenolan Caves Mountain House, Blue Mountains, Australia, November.

Campbell, Nick and Mary Beckman. 1997. Stress, prominence, and spectral tilt. In A. Botinis, G. Kouroupetroglou, and G. Carayiannis, editors, *Intonation: Theory, Models and Applications*, pages 67–70, Athens. ESCA.

Caspers, Johanneke. 1998. Who's next? the melodic marking of question vs. continuation in dutch. *Language and Speech: Special Issue on Prosody and Conversation*, 41(3-4).

Chafe, Wallace. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In C. Li, editor, *Subject and Topic*. The Academic Press, New York, pages 25–55.

Cooper, W. and J. Paccia-Cooper. 1980. *Syntax and Speech*. Harvard University Press, Cambridge MA.

Culicover, Peter W. and Michael Rochemont. 1983. Stress and focus in English. *Language*, 59(1):123–165, March.

Cutler, A. and D. Foss. 1977. On the role of sentence stress in sentence processing. *Language and Speech*, 20:1–10.

Cutler, A and eds. Ladd, D. R. 1983. *Prosody: Models and Measurements*. Springer-Verlag, Berlin.

Dirksen, A. 1992. Accenting and deaccenting: A declarative approach. In *Proceedings of COLING-92*, pages 865–869.

Dirksen, Arthur and Hugo Quene. 1993. Prosodic analysis: The next generation. In Vincent J. van Hueven and Louis C. W. Pols, editors, *Analysis and Synthesis of Speech: Strategic Research towards High-Quality Text-to-Speech Generation*. Mouton de Gruyter, pages 131–144.

Downing, B. 1970. *Syntactic Structure and Phonological Phrasing in English*. Ph.D. thesis, University of Texas, Austin.

Enkvist, N. 1979. Marked focus: Functions and constraints. In S. Greenbaum, G. Leech, and J. Svartvik, editors, *Studies in English Linguistics for Randolph Quirk*. Longmans, London, pages 134–152.

Erteschik-Shir, Nomi and Shalom Lappin. 1983. Under stress: A functional explanation of English sentence stress. *Journal of Linguistics*, 19:419–453.

European Speech Communication Association. 1994. *Conference Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*, Mohonk Mountain House, New Paltz, N. Y. .

Fowler, C. A. and J. Housum. 1987. Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, 26:489–504.

Fuchs, A. 1980. Accented subjects in 'all-new' utterances. In G. Brettschneider and C. Lehmann, editors, *Wege zur Universalienforschung: sprachwissenschaftliche Beitrage zum 60.* Narr, Tubingen.

Fuchs, A. 1984. Deaccenting and default accent. In D. Gibbon and H. Richter, editors, *Intonation, Accent and Rhythm*. Walter de Gruyter, Berlin, pages 134–164.

Fujio, Shigeru, Yoshinori Sagisaka, and Norio Higuchi. 1997. Prediction of major phrase boundary location and pause insertion using a stochastic context-free grammar. In Yoshinori Sagisaka, Nick Campbell, and Norio Higuchi, editors, *Computing Prosody: Computational Models for Processing Spontaneous Speech*. Springer, pages 271–283.

Gawronska, Barbara and David House. 1998. Information extraction and text generation of news reports for a Swedish-English bilingual spoken dialogue system. In *Proceedings of ICSLP-98*, Sydney. International Conference on Spoken Language Processing.

Gee, J. P. and F. Grosjean. 1983. Performance structure: A psycholinguistic and linguistic apprasial. *Cognitive Psychology*, 15:411–458.

Geluykens, Ronald and Marc Swerts. 1994. Prosodic cues to discourse boundaries in experimental dialogues. *Speech Communication*, 15:69–77.

Gorin, Al. 1995. Spoken dialog as a feedback control system. In Paul Dalsgaard, Lars Bo Larsen, Louis Boves, and Ib Thomsen, editors, *Proceedings of the ESCA Workshop on Spoken Dialogue Systems: Theories and Applications*, Visgo.

Grabe, E., C. Gussenhoven, J. Haan, E. Marsi, and B. Post. 1997. The meaning of intonation phrase onsets in Dutch. In A. Botinis, G. Kouroupetroglou, and G. Carayiannis, editors, *Intonation: Theory, Models and Applications*, pages 161–164, Athens. ESCA.

Grosjean, F., L. Grosjean, and H. Lane. 1979. The patterns of silence: Performance structures in sentence production. *Cognitive Psychology*, 11:58–81.

Grosz, Barbara and Julia Hirschberg. 1992. Some intonational characteristics of discourse structure. In *Proceedings of the International Conference on Spoken Language Processing*, Banff, October. ICSLP.

Gundel, J. 1978. Stress, pronominalization, and the given-new distinction. *University of Hawaii Working Papers in Linguistics*, 10(2):1–13.

Gussenhoven, Carlos. 1983. *On the Grammar and Semantics of Sentence Accents*. Foris Publications, Dordrecht.

Gussenhoven, Carols and Toni Rietveld. 1997. Empirical evidence for the contrast between l* and h* in dutch rising contours. In A. Botinis, G. Kouroupetroglou, and G. Carayiannis, editors, *Intonation: Theory, Models and Applications*, pages 18–20, Athens. ESCA.

Hess, Wolfgang, Anton Batliner, Andreas Kiessling, Falf Kompe, Elmar Nöth, Anja Petzold, Matthias Reyelt, and Volker Strom. 1997. Prosodic modules for speech recognition and understanding in VERBMOBIL. In Yoshinori Sagisaka, Nick Campbell, and Norio Higuchi, editors, *Computing Prosody: Computational Models for Processing Spontaneous Speech*. Springer, pages 361–382.

Hirschberg, J. and Gregory Ward. 1992. The influence of pitch range, duration, amplitude, and spectral features on the interpretation of **l*+h l h%**. *Journal of Phonetics*, 20(2):241–251.

Hirschberg, Julia. 1993. Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence*, 63:305–340, October.

Hirschberg, Julia and Cinzia Avesani. 1997. The role of prosody in disambiguating potentially ambiguous utterances in English and Italian. In *ESCA Tutorial and Research Workshop on Intonation: Theory, Models and Applications*, Athens, September.

Hirschberg, Julia, Diane Litman, and Marc Swerts. 1999. Prosodic cues to recognition errors. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU'99).*

Hirschberg, Julia and Christine Nakatani. 1996. A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of the 34th Annual Meeting*, Santa Cruz. Association for Computational Linguistics.

Hirschberg, Julia and Pilar Prieto. 1996. Training intonational phrasing rules automatically for English and Spanish text-to-speech. *Speech Communication*, 18:281–290.

Horne, M. 1985. English sentence stress, grammatical functions and contextual coreference. *Studia Linguistica*, 39:51–66.

Horne, M. 1991a. Accentual patterning in 'new' vs 'given' subjects in English. Working Papers 36, Department of Linguistics, Lund University, Lund.

Horne, M. 1991b. Phonetic correlates of the new/given parameter. In *Proceedings of the Twelfth International Congress of Phonetic Sciences*, pages 230–233, Aix-en-Provence. ICPhS.

Horne, Merle. 1987. Towards a discourse-based model of English sentence intonation. Working Papers 32, Lund University Department of Linguistics.

Horne, Merle and Marcus Filipsson. 1994. Computational extraction of lexico-grammatical information for generation of Swedish intonation. In

Conference Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis, pages 220–223.

House, David and Paul Touati, editors. 1993. *Working Papers 41: Proceedings of an ESCA Workshop on Prosody*, Lund.

Jackendoff, Ray S. 1972. *Semantic Interpretation in Generative Grammar.* MIT Press, Cambridge MA.

1998. *Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves Mountain House, Blue Mountains, Australia, November.

Klabbers, Esther, Emiel Krahmer, and Mariet Theune. 1998. A generic algorithm for generating spoken monologues. In *Proceedings of ICSLP-98*, Sydney. International Conference on Spoken Language Processing.

Koike, Kazuhito, Hirotaka Suzuki, and Hiroaki Saito. 1998. Prosodic parameters in emotional speech. In *Proceedings of ICSLP-98*, Sydney. International Conference on Spoken Language Processing.

Koiso, Hanae, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu Den. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. *Language and Speech: Special Issue on Prosody and Conversation*, 41(3-4).

Koiso, Hanae, Atsushi Shimojima, and Yasuhiro Katagiri. 1998. Collaborative signaling of informational structures by dynamic speech rate. *Language and Speech: Special Issue on Prosody and Conversation*, 41(3-4).

Kompe, R., E. Nöthe, A. Kiessling, T. Kuhn, M. Mast, H. Niemann, K. Ott, and A. Batliner. 1994. Prosody takes over: Towards a prosodically guided dialog system. *Speech Communication*, 15:155–167.

Krahmer, Emil and Marc Swerts. 1998. Reconciling two competing views on contrastiveness. In *Proceedings of ICSLP-98*, Sydney. International Conference on Spoken Language Processing.

Kruyt, J. G. 1985. *Accents from Speakers to Listeners: An Experimental Study of the Production and Perception of Accent Patterns in Dutch.* Ph.D. thesis, University of Leyden.

Ladd, D. R. 1979b. Light and shadow: A study of the syntax and semantics of sentence accents in English. In L. Waugh and F. van Coetsem, editors, *Contributions to Grammatical Studies: Semantics and Syntax.* University Park Press, Baltimore, pages 93–131.

Ladd, D. Robert. 1977. The function of the a-rise accent in English. Distributed by the Indiana University Linguistics Club.

Ladd, D. Robert. 1978. Stylized intonation. *Language*, 54:517–540.

Ladd, D. Robert. 1980. *The Structure of Intonational Meaning.* Indiana University Press, Bloomington, Ind.

Ladd, D. Robert. 1996. *Intonational phonology.* Cambridge University Press.

Lakoff, George. 1971. Presupposition and relative well-formedness. In *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics, and Psychology.* Cambridge University Press, Cambridge UK, pages 329–340.

Lehiste, I. 1979. Perception of sentence and paragraph boundaries. In B. Lindblom and S. Oehman, editors, *Frontiers of Speech Research*. Academic Press, London, pages 191–201.

Lehman, C. 1977. A re-analysis of givenness: Stress in discourse. In *Papers from the 13th Annual Meeting*, pages 316–. Chicago Linguistic Society.

Levow, Gina-Anne. 1998. Characterizing and recognizing spoken corrections in human-computer dialogue. In *Proceedings of the 36th Annual Meeting of the Association of Computational Linguistics, COLING/ACL 98*, pages 736–742.

Liberman, Mark and Ivan A. Sag. 1974. Prosodic form and discourse function. In *Papers of the Tenth Regional Meeting*, pages 416–427. Chicago Linguistic Society.

Murray, Iain R. and Mohn L. Arnott. 1993. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America*, 93(2):1097–1108.

Nooteboom, S. G. and J. Terken. 1982. What makes speakers omit pitch accents?: An experiment. *Phonetica*, 39:317–336.

Nöth, Elmar, A. Batliner, V. Warnke, J. Haas, M. Boros, J. Buckow, R. Huber, F. Gallwitz, M. Nutt, and H. Niemann. 2000. On the use of prosody in automatic dialogue understanding. *Speech Communication.* This volume.

Ostendorf, M. and N. Veilleux. 1994. A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Computational Linguistics*, 20(1):27–54.

Oviatt, S. L., G. Levow, M. MacEarchern, and K. Kuhn. 1996. Modeling hyperarticulate speech during human-computer error resolution. In *Proceedings of ICSLP-96*, pages 801–804, Philadelphia.

Passoneau, R. and D. Litman. 1993. Feasibility of automated discourse segmentation. In *Proceedings of the 31st Annual Meeting*, Ohio State University. Association for Computational Linguistics.

Pierrehumbert, Janet and Julia Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan, and M. Pollack, editors, *Intentions in Communication*. MIT Press, Cambridge MA, pages 271–311.

Pierrehumbert, Janet B. 1980. *The Phonology and Phonetics of English Intonation*. Ph.D. thesis, Massachusetts Institute of Technology, September. Distributed by the Indiana University Linguistics Club.

Pitrelli, John, Mary Beckman, and Julia Hirschberg. 1994. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings of the Third International Conference on Spoken Language Processing*, volume 2, pages 123–126, Yokohama. ICSLP.

Prevost, Scott and Mark Steedman. 1994. Specifying intonation from context for speech synthesis. *Speech Communication*, 15:139–153.

Price, P. J., M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong. 1990. The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America*, December.

Prince, E.F. 1981. Toward a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*. The Academic Press, New York, pages 223–255.

Rank, Erhard and Hannes Pirker. 1998. Generating emotional speech with a concatenative synthesizer. In *Proceedings of ICSLP-98*, Sydney. International Conference on Spoken Language Processing.

Rochemont, Michael S. and Peter W. Culicover. 1990. *English Focus Constructions and the Theory of Grammar*. Cambridge University Press, Cambridge UK.

Rooth, Mats. 1985. *Association with Focus*. Ph.D. thesis, University of Massachusetts, Amherst MA.

Rooth, Mats. 1991. A theory of focus interpretation. Presented at the Workshop on the Syntax and Semantics of Focus, Third European Summer School in Language, Logic and Information, Universitaet des Saarlandes, Saarbrucken.

Sag, I. A. and M. Y. Liberman¡. 1975. The intonational disambiguation of indirect speech acts. In *Papers from the Eleventh Regional Meeting*. Chicago Linguistic Society.

Sagisaka, Yoshinori, Nick Campbell, and Norio Higuchi, editors. 1997. *Computing Prosody: Computational Models for Processing Spontaneous Speech*. Springer.

Sanderman, Angelien A. and Rene Collier. 1997. Prosodic phrasing and comprehension. *Language and Speech*, 40(4):391–408.

Schmerling, S. 1971. Presupposition and the notion of normal stress. In *Papers from the 7th Regional Meeting*, pages 242–253. Chicago Linguistic Society.

Schmerling, S. 1974. A re-examination of the notion NORMAL STRESS. *Language*, 50:66–73.

Schmerling, S. 1975. Evidence from sentence stress for the notions of topic and comment. In S. Schmerling and R. King, editors, *Texas Linguistics Forum*. University of Texas, Department of Linguistics, pages 135–141.

Schmerling, Susan F. 1976. *Aspects of English Sentence Stress*. University of Texas Press, Austin. Revised 1973 thesis, University of Illinois at Urbana.

Schröder, Marc, Veronique Auberge, and Marie-Agnes Cathiard. 1998. Can we hear smile? In *Proceedings of ICSLP-98*, Sydney. International Conference on Spoken Language Processing.

Selkirk, E. 1984. *Phonology and Syntax*. MIT Press, Cambridge MA.

Shimojima, Atsushi, Yasuhiro Katagiri, Hanae Koiso, and Marc Swerts. 2001. An experimental study on the informational and grounding functions of prosodic features of Japanese echoic responses. *Speech Communication*.

Shriberg, Elizabeth, Rebecca Bates, Andreas Stolcke, Paul Taylor, Dan Jurafsky, Klaus Ries, Noah Coccaro, Rachel Martin, Marie Meteer, and Carol van Ess-Dykema. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech: Special Issue on Prosody and Conversation*, 41(3-4).

Silverman, K. 1987. *The Structure and Processing of Fundamental Frequency Contours.* Ph.D. thesis, Cambridge University, Cambridge UK.

Silverman, K., M. Beckman, J. Pierrehumbert, M. Ostendorf, C. Wightman, P. Price, and J. Hirschberg. 1992. ToBI: A standard scheme for labeling prosody. In *Proceedings of the Second International Conference on Spoken Language Processing*, pages 867–879, Banff, October. ICSLP.

Silverman, Kim, Ashok Kalyanswamy, Julie Silverman, Sara Basson, and Dina Yashchin. 1993. Synthesiser intelligibility in the context of a name-and-address information service. In *Proceedings of the 3rd European Conference on Speech Communication and Technology*, volume 3, pages 2169–2172, Berlin. EUROSPEECH-93.

Sproat, Richard, editor. 1998. *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach.* Kluwer.

Swerts, M. 1997. Prosodic features at discourse boundaries of different strength. *Journal of the Acoustical Society of America*, 22(1):25–41.

Swerts, M. and M. Ostendorf. 1997. Prosodic and lexical indications of discourse structure in human-machine interactions. *Speech Communication*, 22:25–41.

Swerts, Marc, Rene Collier, and Jacques Terken. 1994. Prosodic predictors of discourse finality in spontaneous monologues. *Speech Communication*, 15:79–90.

Tamoto, Masafumi and Takeshi Kawabata. 1998. A schema for illocutionary act identification with prosodic features. In *Proceedings of ICSLP-98*, Sydney. International Conference on Spoken Language Processing.

Taylor, Paul, Simon King, Stephen Isard, and Helen Wright. 1998. Intonation and dialog context as constraints for speech recognition. *Language and Speech: Special Issue on Prosody and Conversation*, 41(3-4).

Terken, J. 1984. The distribution of pitch accents in instructions as a function of discourse structure. *Language and Speech*, 27:269–289.

Terken, J. and S. G. Nooteboom. 1987. Opposite effects of accentuation and deaccentuation on verification latencies for given and new information. *Language and Cognitive Processes*, 2(3/4):145–163.

Terken, J. M. B. 1985. *Use and Function of Accentuation: Some Experiments.* Ph.D. thesis, University of Leiden, Helmond, Neth.

Terken, Jacques. 1997. Variation of accent prominence within the phrase: Models and spontaneous speech data. In Yoshinori Sagisaka, Nick Campbell, and Norio Higuchi, editors, *Computing Prosody: Computational Models for Processing Spontaneous Speech.* Springer, pages 95–116.

Terken, Jacques and Julia Hirschberg. 1994. Deaccentuation of words representing 'given' information: Effects of persistence of grammatical function and surface position. *Language and Speech*, 37(2):125–145.

van Donzel, Monique. 1999. *Prosodic Aspects of Information Structure in Discourse.* Holland Academic Graphics.

van Hueven, Vincent J. and Louis C. W. Pols, editors. 1993. *Analysis and*

Synthesis of Speech: Strategic Research towards High-Quality Text-to-Speech Generation. Mouton de Gruyter.

van Santen, Jan P. H., Richard Sproat, Joseph P. Olive, and Julia Hirschberg, editors. 1997. *Progress in Speech Synthesis.* Springer.

Wade, E., E. E. Shriberg, and P. J. Price. 1992. User behaviors affecting speech recognition. In *Proceedings of ICSLP-92*, volume 2, pages 995–998, Banff.

Wales, Roger and Hugh Toner. 1979. Intonation and ambiguity. In W. E. Cooper and E. C. Walker, editors, *Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett.* Halsted Press, New York.

Ward, G. and J. Hirschberg. 1985. Implicating uncertainty: The pragmatics of fall-rise intonation. *Language*, 61(4):747–776, December.

Warnke, V., R. Kompe, H. Niemann, and E. Nöth. 1997. Integrated dialog act segmentation and classification using prosodic features and language models. In *Proceedings of EUROSPEECH-97*, volume 1, pages 207–210, Rhodes.

Whiteside, Sandra P. 1998. Simulated emotions: an acoustic study of voice and pertubation measures. In *Proceedings of ICSLP-98*, Sydney. International Conference on Spoken Language Processing.

Williams, Sandra. 1998. Generating pitch accents in a concept-to-speech system using a knowledge base. In *Proceedings of ICSLP-98*, Sydney. International Conference on Spoken Language Processing.

Wilson, Dierdre and Dan Sperber. 1979. Ordered entailments: An alternative to presuppositional theories. In C.-K. Oh and D. A. Dinneen, editors, *Syntax and Semantics*, volume 11. The Academic Press, New York, pages 229–324.

Zacharski, Ron. 1992. Generation of accent in nominally premodified noun phrases. In *Papers Presented to the 15th International Conference on Computational Linguistics*, pages 253–259, Nantes. International Conference on Computational Linguistics.

Notes

[1] For a general overview of work on the functions of prosody, work by Dwight Bolinger (Bolinger, 1986; Bolinger, 1989) and Bob Ladd (Ladd, 1980; Ladd, 1996). For a sample of individual research efforts in the general field of intonation studies, see the proceedings of the ESCA workshops on intonation in 1993 and 1997 (House and Touati, 1993; Botinis, Kouroupetroglou, and Carayiannis, 1997). To get a good view of the application of intonational research to text-to-speech systems, see the proceedings of ESCA workshops on text-to-speech (Benoit and Bailly, 1989; Mohonk, 1994; van Santen et al., 1997; Jenolan, 1998) and collected articles in (Cutler and Ladd, 1983; van Hueven and Pols, 1993; Sagisaka, Campbell, and Higuchi, 1997; Sproat, 1998).

[2] A fuller description of the ToBI systems may be found in the ToBI conventions document and the training materials available at http://ling.ohio-state.edu/ tobi.

[3] **H* L-L%** for the first two in American English; **L* H-H%** for the third — meaning that accented items in the phrase generally bear accents of a typical category and phrases end in a typical phrase accent/boundary tone combination.

[4] Accent indicators in the following examples should be interpreted as associated with the words they precede.

[5] Below, boundaries are marked by |.