Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 9946–9952, October 1995
Colloquium Paper

This paper was presented at a colloquium entitled "Human–Machine Communication by Voice," organized by Lawrence R. Rabiner, held by the National Academy of Sciences at The Arnold and Mabel Beckman Center in Irvine, CA, February 8–9, 1993.

Linguistic aspects of speech synthesis

JONATHAN ALLEN

Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139-4307

ABSTRACT **The conversion of text to speech is seen as an analysis of the input text to obtain a common underlying linguistic description, followed by a synthesis of the output speech waveform from this fundamental specification. Hence, the comprehensive linguistic structure serving as the substrate for an utterance must be discovered by analysis from the text. The pronunciation of individual words in unrestricted text is determined by morphological analysis or letter-to-sound conversion, followed by specification of the word-level stress contour. In addition, many text character strings, such as titles, numbers, and acronyms, are abbreviations for normal words, which must be derived. To further refine these pronunciations and to discover the prosodic structure of the utterance, word part of speech must be computed, followed by a phrase-level parsing. From this structure the prosodic structure of the utterance can be determined, which is needed in order to specify the durational framework and fundamental frequency contour of the utterance. In discourse contexts, several factors such as the specification of new and old information, contrast, and pronominal reference can be used to further modify the prosodic specification. When the prosodic correlates have been computed and the segmental sequence is assembled, a complete input suitable for speech synthesis has been determined. Lastly, multilingual systems utilizing rule frameworks are mentioned, and future directions are characterized.**

To facilitate human–machine communication, there is an increasing need for computers to adapt to human users. This means of interaction should be pleasant, easy to learn, and reliable. Since some computer users cannot type or read, the fact that speech is universal in all cultures and is the common basis for linguistic expression means that it is especially well suited as the fabric for communication between humans and computer-based applications. Moreover, in an increasingly computerized society, speech provides a welcome humanizing influence. Dialogues between humans and computers require both the ability to recognize and understand utterances and the means to generate synthetic speech that is intelligible and natural to human listeners. In this paper the process of converting text to speech is considered as the means for converting text-based messages in computer-readable form to synthetic speech. Both text and speech are physically observable surface realizations of language, and many attempts have been made to perform text-to-speech conversion by simply recognizing letter strings that could then be mapped onto intervals of speech. Unfortunately, due to the distributed way in which linguistic information is encoded in speech, it has not been possible to establish a comprehensive system utilizing these correspondences. Instead, it has been necessary to first analyze the text into an underlying abstract linguistic structure that is common to both text and speech surface realizations. Once this structure is obtained, it can be used to drive the speech synthesis process in order to produce the desired output acoustic signal. Thus, text-to-speech conversion is an *analysis-synthesis system*. The analysis phase must detect and describe language patterns that are implicit in the input text and that are built from a set of abstract linguistic objects and a relational system among them. It is inherently linguistic in nature and provides the abstract basis for computing a speech waveform consistent with the constraints of the human vocal apparatus. The nature of this linguistic processing is the focus of this paper, together with its interface to the signal processing composition process that produces the desired speech waveform.

As in many systems, the complexity of the relationship between the text input and the speech output forces levels of intermediate representation. Thus, the overall conversion process is broken up through the utilization of two distinct hierarchies. One of these is *structural* in nature and is concerned with the means to compose bigger constructs from smaller ones (e.g., sentences are composed of words). The second hierarchy is an arrangement of different abstractions that provide *qualitatively differing* constraint domains that interact to characterize all linguistic forms. These abstract domains include phonetics, phonology, the lexicon, morphology, syntax, semantics, acoustics, anatomy, physiology, and computation. In computing the overall text-to-speech process, these hierarchies are exploited to provide the environment for encoding relationships between linguistic entities. In this way, as the linguistic framework is built up, algorithms are utilized to produce additional facts, thus further extending the total characterization of the desired utterance. Thus, words can be "parsed" to discover their constituent morphemes, each of which corresponds to a lexical entry that provides both the phonological and the syntactic nature of the morpheme. The goal of the conversion process is to produce a comprehensive framework sufficient to allow the computation of the output speech waveform. Furthermore, we take as a working hypothesis the proposition that *every aspect of linguistic structure manifests itself in the acoustic waveform*. If this is true, the analysis part of the conversion process must provide a *completely specified* framework in order to ensure that the output speech waveform will be responsive to all linguistic aspects of the utterance.

Given the need to derive this structural framework, we can seek to understand the nature of the framework, how it is represented, how it can be discovered, and how it can be interpreted to produce synthetic speech. Answers to these questions are found from study of the various constraints on speech and language production, to which we turn now.

CONSTRAINTS ON SPEECH PRODUCTION

For any text-to-speech system, the process by which the speech signal is generated is constrained by several factors. The *task*

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

in which the system is used will constrain the number and kind of speech voices required (e.g., male, female, or child voices), the size and nature of the vocabulary and syntax to be used, and the message length needed. Thus, for restricted systems such as those that provide announcements of arrivals and departures at a railroad station, the messages are very short and require only limited vocabulary, syntax, and range of speaking style, so a relatively simple utterance composition system will suffice. In this paper, however, it is assumed that the vocabulary, syntax, and utterance length are *unrestricted* and that the system must strive to imitate a native speaker of the language reading aloud. For the *language* being used, the linguistic structure provides many constraining relationships on the speech signal. The phonetic repertoire of sounds; the structure of syllables, morphemes, words, phrases, and sentences; the intended meaning and emphasis; and the interactive dialogue pattern restrict the class of possible linguistic structures. While many of the techniques described here are applicable to several languages, most of the results cited are for English. (Multilingual systems are described in a later section.) Of course, for all speakers, the *human vocal apparatus* limits the class of signals that can emanate from the lips and nose. The oral and nasal passages serve as a time-varying filter to acoustic disturbances that are excited either by the vocal cords or frication generated at some constriction in the vocal tract. All of these constraints, acting together, drive the speech generation process, and hence the text-to-speech process must algorithmically discover the overall ensemble of constraints to produce the synthetic speech waveform.

WORD-LEVEL ANALYSIS

Early attempts to build text-to-speech systems sought to discover direct *letter-to-sound* relationships between letter strings and phoneme sequences (1). Unfortunately, as noted above, a linguistic analysis is needed, and there is a consequent need for a constraining lexicon. This dictionary is used in several ways. Borrowed foreign words, such as "parfait" and "tortilla" retain their original pronunciation and do not follow the letter-to-sound rules of the language that imports them. Also, closed-class (function) words differ in pronunciation from open-class words. Thus, the letter "f" in "of" is pronounced with vocal cord voicing, whereas the "f" in open-class words such as "roof" is unvoiced. Similarly, the "a" in "have" is pronounced differently than the "a" in "behave" and other open-class words. Hence, it makes sense to place these frequently occurring function words in a lexicon, since otherwise they will needlessly complicate any set of pronunciation rules. If the dictionary contains *morphs* (the surface textual realizations of abstract morphemes) rather than words, then algorithms can be introduced (2, 3) to discover morph boundaries within words that delimit text letter strings that can be used to determine corresponding phoneme sequences. Thus, there are many pronunciations of the letter string "ea" as found in "reach," "tear," "steak," and "leather," but the "ea" in "changeable" is broken up by the internal morph boundary, and hence the "ea" is not functioning as a vowel digraph for purposes of pronunciation. Similarly, the "th" in "dither" is functioning as a consonant cluster, but in "hothouse" there is a morph boundary between the "t" and the "h," thus breaking up the cluster. For all of these reasons, a morph lexicon is both necessary and essential to algorithms that determine the pronunciation of *any* English word. Furthermore, contemporary lexicons "cover" over 99 percent of all words and provide much more accurate pronunciations than letter-to-sound rules, which will be discussed later.

Given a morph lexicon, word-level linguistic analysis consists of finding the constituent morphemes (and morphs) of each word. These units have a number of valuable properties, in addition to those already noted above. Morphemes are the *atomic* minimal syntactic units of a language, and they are very *stable* in the language in that new morphemes are rarely introduced, and existing ones are rarely dropped from the language. These morphemes have large *generative power* to make words, so that a morph lexicon of given size can easily cover at least an order-of-magnitude larger number of words. Furthermore, as we have seen above, many language phenomena extend only within morph boundaries, and regularly inflected words (e.g., "entitled") and regular compound words (e.g., "snowplow") are readily *covered* by lexical morphemes.

The parsing of words to reveal their constituent morphemes (2, 3) is an interesting recursive process that must recognize the mutating effects of vocalic suffixes. There are several such changes, such as consonant doubling to produce "fitted" from "fit + ed," the change of "y" to "i" in "cities" from "city + es," and restoration of the final silent "e" as in "choking" from "choke + ing." Note that in each of these cases the vocalic nature of the first letter of the suffix triggers the mutation that takes place during the morph composition process, and it is this change that must be undone in order to recognize the constituent lexical morphs in a word. In addition to the difficulties introduced by these mutations, it turns out that the parsing of words into morphs is ambiguous, so that, for example, "scarcity" can be covered by "scar + city," "scarce + ity," or "scar + cite + y." Fortunately, a simple test that prefers affixed forms over compounds can accurately pick the correct parse. It is interesting that these tests apply to the abstract morphemic structure of the parse, rather than the surface morph covering. For example, "teething" can be parsed into both "teethe + ing" and "teeth + ing," but in the latter analysis "teeth" is already an inflected form ("tooth" + PLURAL), and the parsing tests will prefer the simpler earlier analysis that contains only one inflection. Comprehensive experience with morphemic analysis, together with the systematic construction of large morph lexicons, have provided a robust basis for computing the pronunciation of individual words, and these techniques are now used in all high-performance text-to-speech systems.

LETTER-TO-SOUND RULES

We have already noted the ability of morph-covering analyses to cover over 99 percent of all words. Consequently, letter-to-sound analysis is attempted only when a morph covering is unavailable, since experience shows that phoneme strings obtained by letter-to-sound analysis are inferior to those found through morph analysis. Since letter-to-sound correspondences do not apply across morph boundaries, any word subjected to letter-to-sound analysis must have any detectable affixes stripped off, leaving a presumed root word for further analysis. Thus, the word "theatricality" is analyzed to "theatr + ic + al + ity." The string of three suffixes is tested for correctness by a compact categorical grammar. Thus, the terminal suffix "ity" produces nouns from adjectives, the medial suffix "al" produces adjectives from nouns or adjectives, and the initial suffix "ic" produces adjectives from nouns. In this way the suffixes are seen to be compatible in terms of their parts-of-speech properties, and hence the string of suffixes is accepted.

Once affixes have been stripped, the residual root is searched for recognizable letter strings that are present in known letter-string-to-phoneme-string correspondences. Consonant clusters are searched for first, since their pronunciation is more stable than vowel clusters, longest string first. Hence, the string "chr" will be found first in "Christmas," while the shorter string "ch" is found in "church." Vowel correspondences are least reliable and are established last in the overall process using both text and the computed phoneme environments. Vowel digraphs are the hardest strings to convert, and "ea" is subject to no fewer than 14 rule environments. Examples include "reach," "tear," "steak," "leather," and "theatricality." A complete algorithm has been described by Allen *et al.* (2).

The advent of large lexicons in machine-readable form, together with modern computing platforms and searching algorithms, have led to sets of letter-to-sound rules that effectively complement the morphemic analysis procedures. Detailed informational analyses (4) have been performed that permit the rational choice of rule correspondences, together with a quantitative assessment of the contribution of each letter or phoneme in the rule context to the accuracy of the rule. For specific applications, desired pronunciations of words, whether they would be analyzed by morph covering or letter-to-sound procedures, can be forced by the simple expedient of placing the entire word directly in the lexicon and hence treating it as an exception. A particularly difficult specific application is the pronunciation of surnames, as found in, say, the Manhattan telephone directory, where many names of foreign origin are found. In this case, etymology is first determined from spelling using trigram statistics (probability estimates of strings of three adjacent letters) (5, 6). Then specialized rules for each language can be utilized. Thus, the "ch" in "Achilles" is pronounced differently than the "ch" in "Church." It is interesting that the use of simple letter statistics, which reflect in part the phonotactics of the underlying language, can be combined with other constraints to produce good results on this exceedingly difficult task, where the frequency distribution of surnames is very different than for ordinary words.

MORPHOPHONEMICS AND LEXICAL STRESS

When morph pronunciations are composed, adjustments take place at their boundaries. Thus, the PLURAL morpheme, normally expressed by the morph "s" or "es," takes on differing pronunciation based on the value of the voicing feature of the root word to which the suffix attaches. Hence, the plural of "horse" requires that a short neutral vowel be inserted between the end of the root and the phonemic realization of PLURAL, else the plural form would only lengthen the terminal /s/ in "horse." On the other hand, if the root word does not end in an s-like phoneme, pronunciation of the plural form has the place and fricative consonant features of /s/ but follows the voicing of the root. Since "dog" ends in a voiced stop consonant, its plural suffix is realized as a /z/, while for the root "cat," terminated by an unvoiced stop consonant, the plural is realized as the unvoiced fricative /s/. A similar analysis applies to the computation of the pronunciation of the morpheme affix PAST, as in "pasted," "bagged," and "plucked." There are additional morphophonemic rules used in text-to-speech systems (2), and they are all highly regular and productive. Without their use, the lexicon would be needlessly enlarged.

One of the major achievements of modern linguistics is the understanding of the lexical stress system of English (7). Prior to the mid-1950s, the stress contours of words were specified by long word lists of similar stress pattern, and those learning English as a second language were expected to assimilate these lists. Over the past 40 years, however, comprehensive rules have been derived whose application computes the surface stress contour of words from an underlying phonological specification. These rules are complex in nature, and apply not only to monomorphemic roots, but also to affixed words and compounds. This elegant theory is remarkably robust and has been extensively tested over large lexicons. The trio of words "system, systematic, systematize" illustrates the substantial stress shifts that can take place not only in the location of stress (first syllable in "system," third syllable in "systematic") but also in stress value (the vowel corresponding to the letter "a" is fully realized and stressed in "systematic" but reduced to a neutral vowel in "systematize"). Furthermore, it becomes clear that the lexicon must contain the nonreduced forms of vowels, since otherwise the correct pronunciation of affixed words may not be computable. Thus, the second vowel of "human" is reduced, but its underlying nonreduced form must be known in the lexicon, since in "humanity" the vowel is not reduced. Some suffixes are never reduced, as "eer" in "engineer," and always receive main stress. The rules for compounds, such as "snowplow" are simple, placing stress on the first morph, but the detection of long multiword compounds is difficult (8, 9) (see the section titled "Prosodic Marking") and typically depends on heuristic principles. As an example, in "white house" the construction is attributive, and stress is placed on the head of the phrase, "house." But in the textually similar phrase "White House," the capital letters serve to denote a specific house, namely the residence of the President of the United States, and hence the phrase denotes a compound noun, with stress on the first word.

After morphophonemic and lexical stress rules are applied, a number of phonological adjustments are made, based on articulatory smoothing. Alveolar (dental ridge) flapped consonants are produced as rapid stops in words such as "butter," and two voiceless stop consonants can assimilate, as in "Pat came home," where the "t" is assimilated into the initial consonant of the word "came." Sometimes the effect of articulatory smoothing must be resisted in the interest of intelligibility, as in the insertion of a glottal stop between the two words "he eats." Without this hiatus mechanism, the two words would run on into one, possibly producing "heats" instead of the desired sequence.

ORTHOGRAPHIC CONVENTIONS

Abbreviations and symbols must be converted to normal words in order for a text-to-speech system to properly pronounce all of the letter strings in a sample of text. While the pronunciation of these is largely a matter of convention, and hence not of immediate linguistic interest, some linguistic analysis is often necessary to pick the appropriate pronunciation when there are ambiguous interpretations. The symbol "I" can be used to designate a pronoun, a letter name, or a Roman numeral, and "Dr." can stand for "Doctor" or "Drive." The string "2/3" can indicate the fraction "two-thirds," "February third," or the phrasal string "two slash three." Numbers and currency pose several problems of interpretation, so that "3.45" can be read "three point four five" or "three dollars and forty-five cents." While these ambiguities can often be resolved by heuristic contextual analysis (including syntactic constraints), some conventions are applied inconsistently. In the analysis of one corpus, the string "I.R.S." appeared (with periods) 22 times, whereas "IRS" appeared 428 times. Lest this pattern seem to predict a rule, "N.Y." was found 209 times, whereas "NY" occurred only 14 times! Recently, comprehensive statistical analyses of large corpora have been completed (6), and decision trees (10) have been constructed automatically from a body of classified examples, once a set of features has been specified. As a result, the quality of conversions of abbreviations to phonemic representation has improved markedly, demonstrating the power of statistical classification and regression analysis.

PART-OF-SPEECH ASSIGNMENT

Much of the linguistic analysis used by text-to-speech systems is done at the word level, as discussed above. But there are many important phonological processes that span multiple word phrases and sentences and even paragraph level or discourse domains. The simplest of these constraints is due to syntactic part of speech. Many words vary with their functioning part of speech, such as "wind, read, use, invalid, and survey." Thus, among these, "use" can be a noun or verb and changes its pronunciation accordingly, and "invalid" can be either a noun or an adjective, where the location of main stress indicates the part of speech. At the single-word level, suffixes have considerable constraining power to predict part of speech, so that "dom" produces nouns, as in "kingdom," and "ness" produces nouns, as in "kindness." But in English, a final "s," functioning as an affix, can form a plural noun or a third-person present-tense singular verb, and every common noun can be used as a verb. To disambiguate these situations and reliably compute the functioning part of speech, a dynamic programming algorithm has been devised (11–14) that assigns parts of speech with very high accuracy. Once again, this algorithm relies on a statistical study of a tagged (marked for part-of-speech) corpus and demonstrates the remarkable capabilities of modern statistical techniques.

PARSING

In addition to determining the functioning part of speech for individual words, modern text-to-speech systems also perform some form of limited syntactic analysis, or parsing. These analyses can be used in many ways. As has already been demonstrated, individual word pronunciations can vary with part of speech. In addition, a parsing analysis can provide the structural basis for the marking of prosodic (or suprasegmental) features such as prominence, juncture, and sentence type (declarative, question, or imperative). The accurate calculation of segment durations and pitch contours requires such prosodic marking based on at least minimal syntactic information, or else the resulting speech will be flat, hard to listen to, and even lacking in intelligibility.

Since the justification for parsing is to help provide the structural basis for intelligible and natural-sounding synthetic speech, it has long been assumed that there is a direct relationship between syntactic structure and prosodic structure (the way in which speakers naturally group words). Over the past decade, however, this view has been increasingly challenged, and many phonologists now believe that there are substantial differences between the two structures (15). Nevertheless, the local phrase-level parsing used by contemporary systems provides an initial structure that is very useful, even though it may later be modified to provide the substrate for prosodic marking (next section). An even stronger view would claim that what is desired is the relationship between pragmatic and semantic structure and sound and that any correspondence between syntax and intonation is largely the by-product of the relations between syntax and intonation, on the one hand, and the higher-level constraints of semantics and pragmatics, on the other hand (16). Nevertheless, despite this possibility, phrase-level parsing must for the present provide the needed structural basis given the lack of such higher-level constraints when the system input consists of text alone. When the input is obtained from a message-producing system, where semantic and pragmatic considerations guide the message composition process, alternative prosodic structures may be available for the determination of prosodic correlates (17).

Unfortunately, full clause-level parsing of unrestricted text is an unsolved problem. Nevertheless, phrase-level parsing is fast and reliable and avoids the substantial complexities of clause-level analysis. The main reason for the success of phrase-level analysis is the high syntactic constraining power of determiner sequences in noun phrases and auxiliary sequences in verb phrases. Many text-to-speech systems provide rapid and accurate phrase-level parsing (2, 11) that provides a sufficient base for the instantiation of prosodic cues. Thus, in the classic sentence "He saw the man in the park with the telescope," where determination of the attachment of the prepositional phrases is several-ways ambiguous, the pronunciation (including prosodics) can be derived from the unambiguous phrasal analysis, without the need for resolving the clause-level ambiguity. Of course, clause-level parsing can often be exploited when available, so that a parser for such structures would be useful, providing it failed gracefully to the phrase level when an unambiguous clause-level analysis could not be obtained. Even if such a comprehensive parser were available, however, many researchers do not believe that its benefits outweigh its cost in terms of both computational expense and necessity as input for prosodic algorithms (18), so there is little motivation to extend the scope of syntactic analysis to the clause level.

PROSODIC MARKING

Once a syntactic analysis is determined, it remains to mark the text for prosodic features. These include mainly intonation, prominence, juncture, and sentence type. Speech synthesis procedures can then interpret the segmental phonetic content of the utterance, along with these prosodic markers, to produce the timing and pitch framework of the utterance, together with the detailed segmental synthesis. Many linguistic effects contribute to the determination of these prosodic features. At the lexical level, some words are inherently stressed. For example, in "Hillary might not make cookies for me," the past tense modal auxiliary "might" and the negative "not" express doubt and negation and are reliably stressed, so they are marked for prominence (19). Pronominal reference can also be designated prosodically. Thus, in "She slapped him in the face and then she hit the man," if "him" and "the man" are coreferential, then "hit" receives prominence and "the man" is reduced. But if "him" and "the man" refer to distinct individuals, then "man" is prominent and "hit" is reduced. Correct determination of pronominal reference is not available from simple phrase-level syntactic analysis and must rely on larger scope discourse analysis, which is beginning to be used in text-to-speech systems, but the existence of these phenomena shows the need for, and utility of, such structural information.

As noted in the previous section on parsing, there has been an increasing emphasis during the past decade on prosodic structure (the natural grouping of words in an utterance) as distinct from syntactic structure. The relationship between these two structures was examined linguistically in Selkirk (15), and an emphasis on "performance structures" as natural groupings was presented in Gee and Grosjean (20). These studies emphasized that performance structures have relatively small basic units, a natural hierarchy, and that the resulting overall structure was more balanced than that provided by syntactic constituent analysis. The "performance" aspect of these analyses utilized subjective appraisal of junctural breaks and discovered that word length and the syntactic label of structural nodes played an important role. These natural groupings were found to provide a flatter hierarchical structure than that provided by a syntactic analysis, so that a long verb phrase, "has been avidly reading about the latest rumors in Argentina," which would result in a hierarchy of seven levels in a typical syntactic analysis, would be grouped into three performance chunks—"has been avidly reading" "about the latest rumors" "in Argentina"—which utilize only four levels of hierarchy. The smallest chunks encode what is probably the smallest bundle of coherent semantic information, as suggested by a "case" type of analysis, which is usually associated with more inflected languages than English. That is, the elements of these chunks form a tightly bound package of conceptual information that might be lexicalized into a single lexical item in another language. These chunks are centered on noun or verb heads, such as "in the house" and "will have been reading it," where there are negligible prosodic breaks between the words.

Although the performance structure analysis presented by Gee and Grosjean (20) presupposed a complete syntactic analysis in order to determine the performance structure, Bachenko and Fitzpatrick (18) rejected the need for clausal structure and predicate-argument relations. Furthermore, "readjustment rules" that had been proposed to convert the syntactic structure to that needed for prosodics were abandoned, and an algorithm was provided to generate prosodic phrases that was claimed to be "discourse neutral," with only 14 percent of the phrases studied discourse determined. In this approach there was no need to recognize verb phrase and sentential constituents: only noun phrases, prepositional phrases, and adjectival phrases "count" in the derivation of the prosodic chunks. This was an extremely encouraging result, since a phrase-level parser, of the type described in the previous section, can provide the needed units. Constituency, adjacency, and length were found to be the main factors determining (discourse-neutral) prosodic phrasing. Of course, these prosodic boundaries can be shifted by discourse phrasing, occasioned by emphasis, contrast, parallelism, and coreference, but the phrasing required for a neutral reading can be directly obtained using these phrasal analyses.

Following the analysis introduced by Bachenko and Fitzpatrick (18), there has been much interest in automatically computing prosodic phrase boundaries. Rather than formulating these techniques in terms of rules, statistical techniques have been exploited. Wang and Hirschberg (21) used classification and regression tree (CART) techniques (10) to combine many factors that can affect the determination of these boundaries. In Ostendorf and Veilleux (22) a hierarchical stochastic model of prosodic phrases is introduced and trained using "break index" data. For predicting prosodic phrase breaks from text, a dynamic programming algorithm is provided for finding the maximum probability prosodic parse. These recent studies are very encouraging, as they provide promising techniques for obtaining the prosodic phrasing of sentences based on input text. The evolution of these techniques, from initial linguistic investigations, through psycholinguistic experiments, to the present computational linguistics studies, is extremely interesting, and the interested reader can gain much insight and understanding from the trajectory of references cited. A useful summary is also provided by Wightman *et al.* (23).

Multiword compounds are often hard to analyze, but their perception is highly facilitated with proper prosodic marking. "Government tobacco price support system" and "power generating station control room complex" are two examples of long compounds in need of prosodic cues to reveal their structure. Many of these examples appear to require semantic analysis that is not available, but surprising improvements have been obtained through careful study of many examples (8, 9). In addition to use of the compound stress rule, which places stress for a two-constituent compound (e.g., "sticky bun") on the left, words can be lexically typed in a way that facilitates prediction of stress. Thus measure words (e.g., "pint," "dollar") can combine with a phrase on their right to form a larger phrase that normally takes stress on the right element, as in "dollar bill" and "pint jug." While these rules are useful, for large compound nominals, further heuristics must be applied in addition to the recursive use of the simple compound rule. A rhythm rule can be used to prevent clashes between strong stresses. In this way the stress on "Hall" in "City Hall parking lot" is reduced and that of "City" is raised, so that while "parking" retains the main stress, the next largest stress is two words away, "City." An interesting example of the power of statistics is the use of mutual information (8) to resolve possible ambiguous parsings. For example, "Wall Street Journal" could be parsed as either ([Wall Street] Journal), or (Wall [Street Journal]), where the latter parse would incorrectly imply main stress on "Street." But in a corpus derived from the Associated Press Newswire for 1988, "Wall Street" occurs 636 times outside the context of "Wall Street Journal," whereas "Street Journal" occurs only five times outside this context, and hence the mutual information measure will favor the first (correct) parse and corresponding main stress on "Journal," with "Wall Street" treated as a normal two-word compound.

Of course, virtually any word in a sentence can be emphasized, and if this is marked in the text by underlining or italics, then prominence can be provided for that word.

Lastly, junctural cues are an important aid to perception. In "The dog Bill bought bit him," the reduced relative clause is not explicitly marked, so that a junctural pause after "bought" can indicate to the listener the end of this embedded clause. It has recently been shown (24) that, for a variety of syntactic classes, naive listeners can reliably separate meanings on the basis of differences in prosodic information. These results were obtained from listener judgments of read speech where the ambiguous material was embedded in a larger context. For example, the sentence "They rose early in May." can be used in the following two ways:

- "In spring there was always more work to do on the farm. May was the hardest month. *They rose early in May*."
- "Bears sleep all winter long, usually coming out of hibernation in late April, but this year they were a little slow. *They rose early in May*."

The fact that listeners can often successfully disambiguate sentences from prosodic cues can be used to build algorithms to pick one of several possible parses based on these cues (25). Using the notion of "break index" (a measure of the junctural separation between two neighboring words) introduced by Price *et al.* (24) and statistical training procedures, the candidate parsed text versions are analyzed in terms of these break indices automatically in order to synthesize predicted prosodic structures, which are then compared with the analyzed prosodic structure obtained from the spoken utterance. This is a good example of analysis-by-synthesis processing, where the correct structural version of an utterance is found by synthesizing all possible versions prosodically (at an abstract level of prosodic structure using break indices) and then comparing them with the prosodically analyzed spoken version. While prosodic correlates (e.g., pitch and durations) can rarely be used directly in a bottom-up manner to infer structure, analysis-by-synthesis techniques utilize a scoring of top-down-generated structures to determine by verification the most likely parse.

Standards for the prosodic marking of speech are currently being developed, together with text-based algorithms to create this encoding. Once a large corpus of text is analyzed in this way, and compared with manually provided markings, a rich new enhancement to the overall framework of linguistic analysis will be available, contributing greatly to increased naturalness and intelligibility of synthetic speech.

In this section, emphasis has been placed on determination of prosodic structure, including prosodic boundaries. Once this structure is available, prosodic correlates must be specified. These include durations and the overall timing framework, and the fundamental frequency contour reflecting the overall intonation contour and local pitch accents to mark stress. Not surprisingly, statistical techniques have been developed for the fitting of segmental durations within syllables and higher-level units (26–28). The placement of pitch accent has also been determined by use of classification and regression tree analysis, basing the result on factors such as part of speech of the word and its adjacent words and its position in a larger prosodic constituent (29). These techniques are also able to introduce shifting of accent to avoid rhythmic clash with a stressed syllable in the next word. Once the overall intonational contour and pitch accent determination is made, the corresponding specification can be used as input to an algorithm (30), which will generate the needed fundamental frequency contour.

DISCOURSE-LEVEL EFFECTS

Beyond the sentence level, there are numerous attributes of the overall discourse (31, 32) that influence the prosodic structure of the extended utterance. Since a topic is usually established in a discourse and then comments are made about the topic, facts that were previously established are given reduced prominence when later repeated in the discourse. In the question-answer sequence "What did Joe buy at the mall? Joe bought a boom box at the mall," only "boom box" will receive prominence in the second sentence, since all other syntactic arguments for the verb "buy" have been established in the previous sentence. This phenomenon is called "new/old information," constraining new information to receive prominence, and old information to be reduced. Determination of what is new and what is old is by no means simple, but a variety of counting techniques have been introduced to heuristically estimate the occurrence of old information, all other terms assumed to be new.

There are a variety of focus-shifting transformations available in English to lead the listener to the intended focus of the sentence or perhaps to distract the listener from the normal focus of the sentence. The passive transformation is probably the most frequently occurring example of this effect. Thus, "John bought the books" can be passivized to "The books were bought by John," which can optionally have the agent (John) deleted to form "The books were bought." In this way the initial focus on "John" is first shifted to "the books," and then "John" disappears altogether. In "The likelihood of a tax on the middle class is small, the President thinks," the agent (the President) has been moved to the end of the sentence, with reduced prominence, hence removing the focus of the sentence from him. Such transformations are frequently used to achieve the desired focus, and it is important to mark the sentential prominences accordingly.

Pragmatic knowledge of the world can provide a bias that sometimes overwhelms other constraints. Thus, "He hit the man with the book" is ambiguous. Either "He" hit "the man with the book," or "He hit the man" "with the book." The plausibility of each interpretation can often be inferred from the discourse context and the prosodic structure appropriately marked. It is probably of some help to text-to-speech systems that the reading with the largest pragmatic bias is likely to be perceived, even if the prosodic correlates mark the alternate reading (33). This effect also indicates that a pragmatically rare interpretation (and hence one with substantial new information) must be strongly marked prosodically in order for the intended reading to be perceived.

Discourse-level context may also facilitate the prosodic marking of complex nominals, designate prepositional phrase attachment, and disambiguate conjoined sentences. Thus, in "The bright students and their mentors . . .," "bright" can modify just the "students" or both the "students and their mentors." While there can be no doubt that marking of intended syntactic structure is useful for perception, many of these constructions are inherently ambiguous, and no general techniques are available for the exploitation of discourse structure for purposes of disambiguation. Indeed, relatively little is known concerning discourse structure and how it can be discovered, given only text as input. On the other hand, when synthesis is performed from an abstract message concept (17), rather than only the resultant surface text, discourse structure may be readily available at the abstract level, and hence directly utilized by prosodic marking procedures.

As the development of discourse theory grows, a number of investigators are creating algorithms for the control of prosodics based on discourse constraints. Within a computer-aided instruction application, discourse effects on phrasing, pitch range, accent location, and tune have been demonstrated by Hirschberg and Pierrehumbert (34). In Hirschberg (35), limited discourse-level information, including given/new distinctions, and some information on focus, topic, and contrast, together with refined parts-of-speech distinctions, have been used to assign intonational features for unrestricted text. Discourse connectivity is often signaled with cue phrases, such as "but," "now," "by the way," and "in any case," and their relation to intonation has been described by Hirschberg and Litman (36). For applications where a message is composed using a discourse formalism, new algorithms are likely that will provide more natural synthetic speech, but when unrestricted text is the input and the discourse domain is similarly unrestricted, the determination of contrast, coreference, and new/old information is very difficult, making incorporation of corresponding intonational effects unlikely. Nevertheless, as discourse theory evolves, its relation to prosody will be established, even if it remains difficult to determine the discourse structure from the input provided. Furthermore, a well-developed theory will facilitate input analysis, and the design of applications that can take advantage of the discourse/prosody mappings that become understood. Although much remains to be discovered, the relation of meaning to intonational contours in discourse (37) is of great importance, and the prospect of a system where specific facets of discourse meaning can be manipulated prosodically is indeed very exciting.

MULTILINGUAL SYNTHESIS

Several groups have developed integrated rule frameworks and languages for their design and manipulation (38–40) Using these structures, a flexible formalism is available for expressing rules and for utilizing these rules to "fill in" the coordinated comprehensive linguistic description of an utterance. The complex data structures provided in these systems also facilitate the alignment of constraints across several domains (or levels of representation), such as a textual character string, the names of words, their constituent morphs and phonemes, and the overall syntactic structure. These unified procedures are a considerable improvement over isolated ad hoc rule systems that apply at only one level of linguistic representation. Furthermore, they facilitate the writing of new rules and experimentation with an overall integrated rule system. Thus, it is no surprise that these groups have built several text-to-speech systems for many different languages. Although good-quality text-to-speech systems have resulted from the exploitation of these frameworks, single-language ad hoc systems currently provide better-quality speech, but this state of affairs probably reflects mostly the amount of time spent on refinement of the rules, rather than any intrinsic limitation of the coordinated framework and rule approach.

THE FUTURE

Contemporary text-to-speech systems are available commercially and are certainly acceptable in many applications. There is, however, both much room for improvement and the need for enhancements to increase the intelligibility, naturalness, and ease of listening for the resultant synthetic speech. In recent years much progress has followed from the massive analysis of data from large corpora. Modern classification and decision tree techniques (10) have produced remarkable results where no linguistic theory was available as a basis for rules. In general, the use of standard algorithmic procedures, together with statistical parameter fitting, has been very successful. To further this process, large tagged data bases are needed, using standard techniques that can be employed by many diverse investigators. Such data bases are just beginning to be developed for prosodic phenomena (41), but they can also be extremely useful for enhancing naturalness at the segmental level. While these statistical techniques can often extract a great deal of useful information from both texts and tagged phonetic transcriptions, the quest for appropriate linguistic models must be aggressively extended at all levels of representation. Where good models are available, such as for morphemic structure and lexical stress, the results are exceedingly robust. Linguistic descriptions of discourse are much needed, and a more detailed and principled prosodic theory that could guide both analysis and synthesis algorithms would be exceedingly useful. Of course, for some tasks, such as the conversion of abbreviations and standard symbols, there is relatively little linguistic content, and statistical techniques will have to bear the brunt of the task.

The use of articulation as a basis for phonology (42) and synthesis may provide a fundamental solution to many problems of speech naturalness and may also introduce useful constraints for speech recognition. Many facts of speech production are best represented at the articulatory level, and a rule system focused on articulatory gestures is likely to be simpler than the current rule systems based on acoustic phonetics. Unfortunately, the acquisition of articulatory information is exceedingly difficult, since it involves careful observation of the entire set of speech articulators, many of which are either hidden from normal view or are difficult to observe without perturbing the normal speech production process. Nevertheless, improvements in the understanding and representation of articulation cannot help but improve synthesis, and it is important that the research community make a long-term commitment to the acquisition of this knowledge.

Lastly, contemporary research is benefiting from quickly evolving computational and experimental technology, which will provide the substrate for many new studies, as well as cost-effective systems for many applications. These facilities allow an attack on text and speech analysis at a level of complexity that was not hitherto possible. Future research will utilize statistical discovery procedures to suggest new linguistic formalisms and to organize observations of very large corpora. It is clear that the text-to-speech research community is now positioned to make large improvements in speech quality over extensive texts and also to contribute directly to the overall base of knowledge in linguistics, computational linguistics, phonetics, and articulation.

1. Venezky, R. L. (1970) *The Structure of English Orthography* (Mouton, The Hague, The Netherlands).
2. Allen, J., Hunnicutt, M. S. & Klatt, D. H. (1987) *From Text to Speech: The MITalk System* (Cambridge Univ. Press, London).
3. Allen, J. (1992) in *Advances in Speech Signal Processing*, eds. Furui, S. & Sondhi, M. (Dekker, New York), pp. 741–790.
4. Lucassen, J. M. & Mercer, R. L. (1984) in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 42.5.1–42.5.4.
5. Church, K. W. (1986) in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 2423–2426.
6. Liberman, M. Y. & Church, K. W. (1992) in *Advances in Speech Signal Processing*, eds. Furui, S. & Sondhi, M. (Dekker, New York), pp. 791–831.
7. Chomsky, A. N. & Halle, M. (1968) *Sound Pattern of English* (Harper & Row, New York).
8. Sproat, R. W. (1990) in *Proceedings of the ESCA Workshop on Speech Synthesis*, pp. 129–132.
9. Sproat, R. W. & Liberman, M. Y. (1987) in *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pp. 140–146.
10. Brieman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984) *Classification and Regression Trees* (Wadsworth & Brooks, Monterey, CA).
11. Church, K. W. (1988) in *Proceedings of the Second Conference on Applied Natural Language Processing* (Austin, TX), pp. 136–143.
12. DeRose, S. (1988) *Comput. Linguist.* **14.**
13. Jelinek, F. (1990) in *Readings in Speech Recognition*, eds. Waibel, A. & Lee, K. (Kaufmann, San Mateo, CA).
14. Kupiec, J. (1992) *Comput. Speech Lang.* **6,** 225–242.
15. Selkirk, E. O. (1984) *Phonology and Syntax: The Relation Between Sound and Structure* (MIT Press, Cambridge, MA).
16. Monaghan, A. I. C. (1989) University of Edinburgh Department of Linguistics, Work in Progress 22.
17. Young, S. J. & Fallside, F. (1979) *J. Acoust. Soc. Am.* **66,** 685–695.
18. Bachenko, J. & Fitzpatrick, E. (1990) *Comput. Linguist.* **16,** 155–170.
19. O'Shaughnessy, D. & Allen, J. (1983) *J. Acoust. Soc. Am.* **74,** 1155–1171.
20. Gee, J. P. & Grosjean, F. (1983) *Cognit. Psychol.* **15,** 411–458.
21. Wang, M. Q. & Hirschberg, J. (1992) *Comput. Speech Lang.* **6,** 175–196.
22. Ostendorf, M. & Veilleux, N. M. (1993) *Comput. Linguist.* **19.**
23. Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M. & Price, P. (1992) *J. Acoust. Soc. Am.* **91,** 1707–1717.
24. Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S. & Fong, C. (1991) *J. Acoust. Soc. Am.* **90,** 2956–2970.
25. Ostendorf, M., Wightman, C. W. & Veilleux, N. M. (1993) *Comput. Speech Lang.* **7,** 193–210.
26. Campbell, W. N. (1992) in *Talking Machines: Theories, Models, and Designs*, eds. Bailly, G., Benoit, C. & Sawallis, T. R. (Elsevier, New York), pp. 211–224.
27. Riley, M. D. (1992) in *Talking Machines: Theories, Models, and Designs*, eds. Bailly, G., Benoit, C. & Sawallis, T. R. (Elsevier, New York), pp. 265–273.
28. Van Santen, J. P. H. (1992) in *Talking Machines: Theories, Models, and Designs*, eds. Bailly, G., Benoit, C. & Sawallis, T. R. (Elsevier, New York), pp. 275–285.
29. Ross, K., Ostendorf, M. & Shattuck-Hufnagel, S. (1992) in *Proceedings of the International Conference on Spoken Language Processing*, pp. 365–368.
30. Pierrehumbert, J. B. (1981) *J. Acoust. Soc. Am.* **70,** 985–995.
31. Grosz, B. J., Pollack, M. E. & Sidner, C. L. (1989) in *Foundations of Cognitive Science*, ed. Posner, M. (MIT Press, Cambridge, MA), Chapt. 11.
32. Grosz, B. J. & Sidner, C. L. (1986) *Comput. Linguist.* **12,** 175–204.
33. Wales, R. & Toner, H. (1979) in *Sentence Processing*, eds. Cooper, W. C. & Walker, E. C. T. (Erlbaum, Hillsdale, NJ).
34. Hirschberg, J. & Pierrehumbert, J. B. (1986) in *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pp. 136–144.
35. Hirschberg, J. (1992) in *Talking Machines: Theories, Models, and Designs*, eds. Bailly, G., Benoit, C. & Sawallis, T. R. (Elsevier, New York), pp. 367–376.
36. Hirschberg, J. & Litman, D. (1987) in *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pp. 163–171.
37. Hobbs, J. R. (1990) in *Plans and Intentions in Communication and Discourse*, eds. Cohen, P. R., Morgan, J. & Pollack, M. E. (MIT Press, Cambridge, MA).
38. Carlson, R. & Granstrom, B. (1986) in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 2403–2406.
39. Hertz, S. R. (1990) in *Proceedings of the ESCA Workshop on Speech Synthesis*, pp. 225–228.
40. Van Leeuwen, H. C. & te Lindert, E. (1993) *Comput. Speech Lang.* **2,** 149–167.
41. Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. & Hirschberg, J. (1992) in *Proceedings of the International Conference on Spoken Language Processing*, pp. 867–870.
42. Browman, C. P. & Goldstein, L. (1989) *Phonology* **6,** 201.