Chapter 5

Spoken Output Technologies

5.1 Overview

Yoshinori Sagisaka

ATR Interpreting Telecommunications Research Laboratories, Tokyo, Japan

5.1.1 A Global View of Synthesis Research

Speech synthesis research predates other forms of speech technology by many years. In the early days of synthesis, research efforts were devoted mainly to simulating human speech production mechanisms, using basic articulatory models based on electro-acoustic theories. Though this modeling is still one of the ultimate goals of synthesis research, advances in computer science have widened the research field to include Text-to-Speech (TtS) processing in which not only human speech generation but also text processing is modeled (Allen, Hunnicutt, et al., 1987). As this modeling is generally done by a set of rules derived, e.g., from phonetic theories and acoustic analyses, the technology is typically referred to as speech synthesis by rule.

Figure 5.1 shows the configuration of a standard TtS system. In such systems, as represented by MITalk (Allen, Hunnicutt, et al., 1987), rule-based synthesis has attained highly intelligible speech quality and can already serve in many practical uses. Ceaseless efforts have improved the quality of rule-based synthetic speech, step by step, by alternating speech characteristics analysis with the development of control rules. However, most of this progress has been system dependent, and remains deeply embedded within system architectures in impenetrable meshes of detailed rules and finely tuned control parameters. As a consequence, the expert knowledge that has been incorporated is not available to be shared commonly and can be very hard to replicate in equivalent systems by other researchers.

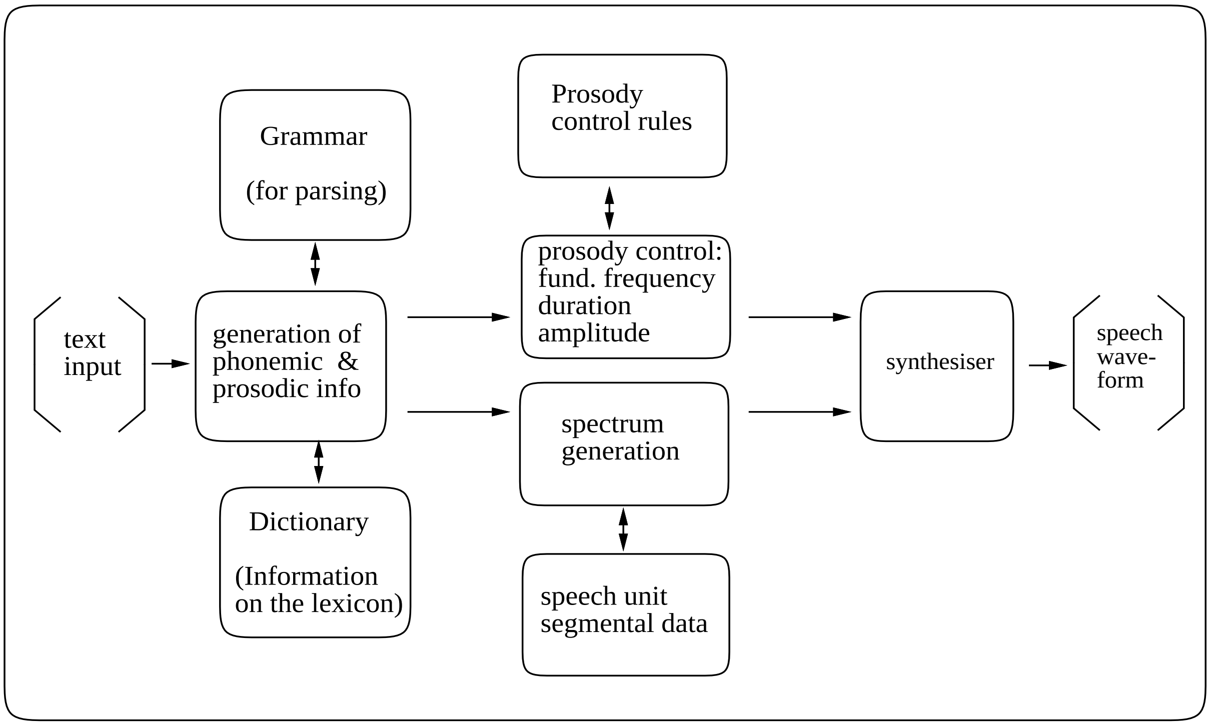

Figure 5.1: The configuration of a standard TtS system.

In contrast to this traditional rule-based approach, a corpus-based approach has also been pursued. In the corpus-based work, well-defined speech data sets have been annotated at various levels with information, such as acoustic-phonetic labels and syntactic bracketing, to serve as the foundation for statistical modeling. Spectral and prosodic feature parameters of the speech data are analyzed in relation to the labeled information, and their control characteristics are quantitatively described. Based on the results of these analyses, a computational model is created and trained using the corpus. By subsequently applying the resulting model to unseen test data, its validity and any defects can be quantitatively shown. By feeding back results from such tests into the original model with extended training, further improvements can be attained in a cyclical process.

As can be easily seen, these formalised procedures characteristic of the corpus-based approach provide for a clear empirical formulation of the controls underlying speech, and with their specific training procedures and their objective evaluation results, can be easily replicated by other researchers with other databases of equivalently annotated speech. In the last decade, the corpus-based approach has been applied to both spectral and prosodic control for speech synthesis. In the following paragraphs, these speech synthesis research activities will be reviewed, with particular emphasis on the types of synthesis unit, on prosody control and on speaker charateristics. Other important topics, such as text processing for synthesis, and spectral parameters and synthesizers, will be detailed in later sections. Through this introduction to the research activities, it will become clear that the corpus-based approach is the key to understanding current research directions in speech synthesis and to predicting the future outcome of synthesis technology.

5.1.2 Synthesis Segment Units

In TtS systems, speech units that are typically smaller than words are used to synthesize speech from arbitrary input text. Since there are over 10,000 different possible syllables in English, much smaller units such as phonemes and dyads (phoneme pairs) have typically been modelled. A speech segment's spectral characteristics vary with its phonetic context, as defined by neighboring phonemes, stress and positional differences, and recent studies have shown that speech quality can be greatly affected by these contextual differences (see for example, Olive, Greenwood, et al., 1993). However, in traditional rule-based synthesis, though these units have been carefully designed to take into account phonetic variations, no systematic studies have been carried out to determine how and where to best extract the acoustic parameters of units, or of what kind of speech corpus can be considered optimal.

To bring objective techniques into the generation of appropriate speech units, unit-selection synthesis has been proposed (Nakajima & Hamada, 1988; Takeda, Abe, et al., 1992; Sagisaka, Kaiki, et al., 1992). These speech units can be automatically determined through the analysis of a speech corpus using a measure of entropy on substrings of phone labels (Sagisaka, Kaiki, et al., 1992). In unit-selection synthesis, speech units are algorithmically extracted from a phonetically transcribed speech data set using objective measures based on acoustic and phonetic criteria. These measures indicate the contextual adequateness of units and the smoothness of the spectral transitions within and between units. Unlike traditional rule-based concatenation synthesis, speech segments are not limited to one token per type, and various types and sizes of units with different contextual variations are used. The phonetic environments of these units and their precise locations are automatically determined through the selection process. Optimal units to match an input phonetic string are then selected from the speech database to generate the target speech output.

The unit selection process involves a combinatorial search over the entire speech corpus, and consequently, fast search algorithms have been developed for this purpose as an integral part of current synthesis. This approach is in contrast to traditional rule-based synthesis where the design of the deterministic units required insights from the researcher's own knowledge and expertise. The incorporation of sophisticated but usually undescribed *knowledge* was the real bottleneck that prevented the automatic construction of synthesis systems.

Corpus-based methods provide for a specification of the speech segments required for concatenative synthesis in three factors:

1. the procedures of the unit selection algorithm;
2. the objective measures used in the selection criteria; and
3. the design of the speech corpus from which the units are extracted.

This modularization of system building is useful not only in reducing construction effort, but also in allowing precise mathematical specification of the problems and in defining ways to cope with them systematically, by improving the selection algorithms, criteria and data.

5.1.3 Prosody Control

For synthesis of natural-sounding speech, it is essential to control prosody, to ensure appropriate rhythm, tempo, accent, intonation and stress. Segmental duration control is needed to model temporal characteristics just as fundamental frequency control is needed for tonal characteristics. In contrast to the relative sparsity of work on speech unit generation, many quantitative analyses have been carried out for prosody control. Specifically, quantitative analyses and modeling of segmental duration control have been carried out for many languages using massive annotated speech corpora (Carlson & Granström, 1986; Bartkova & Sorin, 1987; Klatt, 1987; Umeda, 1975).

Segmental duration is controlled by many language specific and universal factors. In early models, because these control factors were computed independently, through the quantification of control rules, unexpected and serious errors were sometimes seen. These errors were often caused simply by the application of independently derived rules at the same time. To prevent this type of error and to assign more accurate durations, statistical optimization techniques that model the often complex interactions between all the contributing factors have more recently been used.

Traditional statistical techniques such as linear regressive analysis and tree regression analysis have been used for Japanese (Kaiki, Takeda, et al., 1992) and American English (Riley, 1992) respectively. To predict the interactions between syllable and segment level durations for British English a feed-forward neural network has been employed (Campbell, 1992). In this modeling, instead of attempting to predict the absolute duration of segments directly, their deviation from the avarage duration is employed to quantify the lengthening and shortening characteristics statistically. Moreover, hierarchical control has been included by splitting the calculation into the current syllable level and its constituent component levels.

While hierarchical control is desired to simulate human temporal organization mechanisms, it can be difficult to optimize such structural controls globally. Multiple split regression (MSR) uses error minimization at arbitrary hierarchical levels by defining a hierarchical error function (Iwahashi & Sagisaka, 1993). MSR incorporates both linear and tree regressions as special cases and interpolates between them by controlling the tiedness of the control parameters. Additive-multiplicative modeling, too, is also an extension of traditional linear analysis techniques, using bilinear expressions and statistical correlation analyses (Van Santen, 1992). These statistical models can optimize duration control without losing freedom of conditioned exception control.

To generate an appropriate fundamental frequency (F_0) contour when given only text as input, an intermediate prosodic structure needs to be specified. Text processing, as described in section 5.3, is needed to produce this intermediate prosodic structure. F_0 characteristics have been analyzed in relation to prosodic structure by many researchers (Maeda, 1976; Hakoda & Sato, 1980; Pierrehumbert, 1981; Liberman & Pierrehumbert, 1984; Fujisaki, 1992). As with duration control, in early models, F_0 control rules were made only by assembling independently analyzed F_0 characteristics. More recently however, statistical models have been employed to associate F_0 patterns with input linguistic information directly, without requiring estimates of the intermediate prosodic structure (Traber, 1992; Sagisaka, Kaiki, et al., 1992; Yamashita, Tanaka, et al., 1993). In these models, the same mathematical frameworks as used in duration control; i.e., feed-foward neural networks, linear and tree regression models have been used.

These computational models can be evaluated by comparing duration or F_0 values derived from the predictions of the models with actual values measured in the speech corpus for the same test input sentences. Perceptual studies have also been carried out to measure the effect of these acoustical differences on subjective evaluation scores by systematically manipulating the durations (Kato, Tsuzaki, et al., 1992). It is hoped that a systematic series of perceptual studies will reveal more about human sensitivities to the naturalness and intelligibility of synthesized speech scientifically and that time consuming subjective evaluation will no longer be needed.

5.1.4 Speaker Characteristics Control

Speech waveforms contain not only linguistic information but also speaker voice characteristics, as manifested in the glottal waveform of voice excitation and in the global spectral features representing vocal tract characteristics. The glottal waveform has been manipulated using a glottal source model (Fant, Liljencrants, et al., 1985) and female voices (more difficult to model) have been successufuly synthesized. However, it is very difficult to fully automate such parameter extraction procedures and the establishment of an automatic analysis-synthesis scheme is longed for.

As for vocal tract characteristics, spectral conversion methods have been proposed that employ the speaker adaptation technology studied in speech recognition (Abe, Nakamura, et al., 1990; Matsumoto, Maruyama, et al., 1994; Moulines & Sagisaka, 1995). This technology is also a good example of the corpus-based approach. By deciding on a spectral mapping algorithm, a measure for spectral distance and a speech corpora for training of the mapping, non-parametric voice conversion is defined. The mapping accuracy can be measured using the spectral distortion measures commonly used in speech coding and recognition.

5.1.5 Future Directions

As indicated in the above paragraphs, speech synthesis will be studied continuously, aiming all the while at more natural and intelligible speech. It is quite certain that TtS technology will create new speech output applications associated with the improvement of speech quality. To accelerate this improvement, it is necessary to pursue research on speech synthesis in such a way that each step forward can be evaluated objectively and can be shared among researchers. To this end, a large amount of commonly available data is indispensable, and objective evaluation methods should be pursued in relation to perceptual studies. An important issue of concern to speech synthesis technology is the variability of output speech. As illustrated by recent advances in speaker characteristics control, the adaptation of vocal characteristics is one dimension of such variability. We also have to consider variabilities resulting from human factors, such as speaking purpose, utterance situation and the speaker's mental states. These paralinguistic factors cause changes in speaking styles reflected in a change of both voice quality and prosody. The investigation of these variations will contribute to elaborate synthetic speech quality and widen its application fields.

Such progress is not only restricted to TtS technology; future technologies related to the furtherance of human capabilities are also being developed. Human capabilities such as the acquisition of spoken language bear strong relations to the knowledge acquisition used in developing speech synthesis systems. Useful language training tools and educational devices can therefore be expected to come out of the pursuit and modeling of such knowledge acquisition processes. The corpus-based approach is well suited to this purpose, and inductive learning from speech corpora will give us hints on the directions this research must take. To pursue these new possiblities, it is essential for speech synthesis researchers to collaborate with researchers in other fields related to spoken language, and to freshly introduce the methodologies and knowledge acquired in those encounters.

5.2 Synthetic Speech Generation

Christophe d'Alessandro & Jean-Sylvain Liénard

LIMSI-CNRS, Orsay, France

Speech generation is the process which allows the transformation of a string of phonetic and prosodic symbols into a synthetic speech signal. The quality of the result is a function of the quality of the string, as well as of the quality of the generation process itself. For a review of speech generation in English the reader is referred to Flanagan and Rabiner (1973) and Klatt (1987). Recent developments can be found in Bailly and Benoît (1992), and in Van Santen, Sproat, et al. (1995).

Let us examine first what is requested today from a text-to-speech (TtS) system. Usually two quality criteria are proposed. The first one is intelligibility, which can be measured by taking into account several kinds of units (phonemes, syllables, words, phrases). The second one, more difficult to define, is often labeled as *pleasantness* or *naturalness*. Actually the concept of naturalness may be related to the concept of realism in the field of image synthesis: the goal is not to restitute the reality but to suggest it. Thus, listening to a synthetic voice must allow the listener to attribute this voice to some *pseudo-speaker* and to perceive some kind of expressivity as well as some indices characterizing the speaking style and the particular situation of elocution. For this purpose the corresponding extra-linguistic information must be supplied to the system (Granström & Nord, 1992).

Most of the present TtS systems produce an acceptable level of intelligibility, but the naturalness dimension, the ability to control expressivity, speech style and pseudo-speaker identity still are poorly mastered. Let us mention however that users demands vary to a large extent according to the field of application: general public applications such as telephonic information retrieval need maximal realism and naturalness, whereas some applications involving professionals (process or vehicle control) or highly motivated persons (visually impaired, applications in hostile environments) demand intelligibility with the highest priority.

5.2.1 Input to the Speech Generation Component

The input string to the speech generation component is basically a phonemic string resulting from the grapheme to phoneme converter. It is usually enriched with a series of prosodic marks denoting the accents and pauses. With few exceptions the phoneme set of a given language is well defined; thus the symbols are not ambiguous. However the transcript may represent either a sequence of abstract linguistic units (phonemes) or a sequence of acoustic-phonetic units (phones or transitional segments). In the former case (phonological or normative transcript) it may be necessary to apply some transformations to obtain the acoustical transcript. In order to make this distinction clearer let us take a simple example in French. The word "médecin" (medical doctor) may appear in a pronunciation dictionary as "mé–de–cin" /me–dœ–s$\tilde{\varepsilon}$/, which is perfectly correct. But when embedded in a sentence it is usually pronounced in a different way "mèt–cin" /mεt–s$\tilde{\varepsilon}$/. The tense vowel "é" /e/ is realized as its lax counterpart "è" /ε/, the "e" /œ/ disappears, the three syllables are replaced by only two, and the voicing of the plosive /d/ is neutralized by the presence of the unvoiced /s/ which follows. Without such rules the output of the synthesizer may be intelligible, but it may be altered from the point of view of naturalness. Such transformations are not simple; they imply not only a set of phonological rules, but also some considerations on the speech style, as well as on the supposed socio-geographical origin of the pseudo-speaker, and on the speech rate.

Analogously, the prosodic symbols must be processed differently according to their abstraction level. But the problem is more difficult, because there is no general agreement in the phonetic community on a set of prosodic marks that would have a universal value, even within the framework of a given language. A noticeable exception is the ToBI system, for transcription of English (Pitrelli, Beckman, et al., 1994). Each synthesis system defines its own repertory of prosodic entities and symbols, that can be classified into three categories: phonemic durations, accents and pauses.

5.2.2 Prosody Generation

Usually only the accents and pauses, deduced from the text, are transcribed in the most abstract form of the prosodic string. But this abstract form has to be transformed into a flow of parameters in order to control the synthesizer. The parameters to be computed include the fundamental frequency (F_0), and the duration of each speech segment as well as its intensity and timber. A melodic (or intonational) model and a duration model are needed to implement the prosodic structure computed by the text processing component of the speech synthesizer.

F_0 evolution, often considered the main support of prosody, depends as do the phonemic durations on phonetic, lexical, syntactic and pragmatic factors. Depending on the language under study, the melodic model is built on different levels, generally the word level (word accent) and the sentence or phrase level (phrase accent). The aim of the melodic model is to compute F_0 curves. Three major types of melodic models are currently in use for F_0 generation. The first type of melodic model is production-oriented. It aims at representing the commands governing F_0 generation.

This type of model associates melodic commands with word and phrase accents. The melodic command is either an impulse or a step signal. The F_0 contour is obtained as the response of a smoothing filter to these word and phrase commands (Fujisaki & Kawai, 1988). The second type of melodic model is rooted in perception research (Hart, Collier, et al., 1990). Synthetic F_0 contours are derived from stylized natural F_0 contours. At the synthesis stage, the F_0 curves are obtained by concatenation of melodic movements: F_0 rises, F_0 falls, and flat movements. Automatic procedures for pitch contour stylization have been developed (d'Alessandro & Mertens, 1995). In the last type of melodic model, F_0 curves are implemented as a set of target values, linked by interpolation functions (Pierrehumbert, 1981).

The phonemic durations result from multifold considerations. They are in part determined from the mechanical functioning of the synthesizer when the latter is of articulatory nature, or from the duration of the prerecorded segments in the case of concatenative synthesis. Another part is related to the accent. Another one, reflecting the linguistic function of the word in the sentence, is usually related to the syntactic structure. Finally, the last part is related to the situation and pseudo-speaker's characteristics (speech rate, dialect, stress, etc.).

Two or three levels of rules are generally present in durational models. The first level represents co-intrinsic duration variations (i.e., the modification of segment durations that are due to their neighbors). The second level is the phrase level: modification of durations that are due to prosodic phrasing. Some systems also take into account a third level, the syllabic level (Campbell & Isard, 1991).

The other prosodic parameters (intensity, timber) are usually implicitly fixed from the start. However, some research is devoted to voice quality characterization or differences between male and female voices (Klatt & Klatt, 1990).

One of the most difficult problems in speech to date is prosodic modeling. A large body of problems come from text analysis (see section 5.3). But there is also room for improvement in both melodic and durational models. In natural speech the prosodic parameters interact in a way that is still unknown, in order to supply the listener with prosodic information while keeping the feeling of fluentness. Understanding the interplay of these parameters is today one of the hottest topics for research on speech synthesis. For prosodic generation, a move from rule-based modeling to statistical modeling is noticeable, as in many areas of speech and language technology (Van Santen, 1994).

5.2.3 Speech Signal Generation

The last step for speech output is synthesis of the waveform, according to the segmental and prosodic parameters defined at earlier stages of processing.

Speech signal generators (the *synthesizers*) can be classified into three categories: (1) articulatory synthesizers, (2) formant synthesizers, and (3) concatenative synthesizers. Articulatory synthesizers are physical models based on the detailed description of the physiology of speech production and on the physics of sound generation in the vocal apparatus (Parthasarathy & Coker, 1992). Typical parameters are the position and kinematics of articulators. Then the sound radiated at the mouth is computed according to equations of physics. This type of synthesizer is rather far from applications and marketing because of its cost in terms of computation and the underlying theoretical and practical problems still unsolved.

Formant synthesis is a descriptive acoustic-phonetic approach to synthesis (Allen, Hunnicutt, et al., 1987). Speech generation is not performed by solving equations of physics in the vocal apparatus, but by modeling the main acoustic features of the speech signal (Klatt, 1980; Stevens & Bickley, 1991). The basic acoustic model is the source/filter model. The filter, described by a small set of *formants*, represents *articulation* in speech. It models speech spectra that are representative of the position and movements of articulators. The source represents *phonation*. It models the glottal flow or noise excitation signals. Both source and filter are controlled by a set of phonetic rules (typically several hundred). High-quality rule-based formant synthesizers, including multilingual systems, have been marketed for many years.

Concatenative synthesis is based on speech signal processing of natural speech databases. The segmental database is built to reflect the major phonological features of a language. For instance, its set of phonemes is described in terms of diphone units, representing the phoneme-to-phoneme junctures. Non-uniform units are also used (diphones, syllables, words, etc.). The synthesizer concatenates (coded) speech segments, and performs some signal processing to smooth unit transitions and to match predefined prosodic schemes. Direct pitch-syncronous waveform processing is one of the most simple and popular concatenation synthesis algorithms (Moulines & Charpentier, 1990). Other systems are based on multipulse linear prediction (Atal & Remde, 1982), or harmonic plus noise models (Laroche, Stylianou, et al., 1993; Dutoit & Leich, 1993; Richard & d'Alessandro, 1994). Several high-quality concatenative synthesizers, including multilingual systems, are marketed today.

5.2.4 Trends in Speech Generation

Perceptive assessment lies among the most important aspects of speech synthesis research (Van Bezooijen & Pols, 1990; Van Santen, 1993; Kraft & Portele, 1995). When one works on phonetic rule definition or segment concatenation, a robust and quick assessment methodology is absolutely necessary to improve the system. Besides, it is also necessary in order to compare the systems to each other. As far as speech naturalness is concerned the problem is still almost untouched. Nobody knows what speech naturalness is or more generally what is expected from a synthesis system once its intelligibility is rated sufficiently highly. In order to explore this domain it will be mandatory to cooperate with psychologists and human factors specialists.

Although the recent developments of speech synthesis demonstrated the power of the concatenative approach, it seems that there is much room for improvement:

1. **Choice of Non-uniforms and Multi-scale Units** (see section 5.1.2): What are the best synthesis units? this question is rooted in psycholinguistics, and is a challenging problem to phonology.
2. **Speech Signal Modification:** Signal representation for speech is still an open problem, particularly for manipulation of the excitation.
3. **Voice Conversion:** What are the parameters, phonetic description, methods for characterization of a particular speaker, and conversion of the voice of a speaker into the voice of another speaker (Valbret, Moulines, et al., 1992)?

Accurate physical modeling of speech production is still not mature for technological applications. Nevertheless, as both basic knowledge on speech production and the power of computers increase, articulatory synthesis will help in improving formant-based methods, take advantage of computational physics (fluid dynamics equations for the vocal apparatus), and better mimic the physiology of human speech production.

Synthesis of human voice is not limited to speech synthesis. Since the beginning of speech synthesis research, many workers also paid some attention to the musical aspects of voice and to singing (Sundberg, 1987). Like TtS, synthesis of singing finds its motivations both in science and technology: on the one hand singing analysis and synthesis is a challenging field for scientific research, and on the other hand, it can serve for music production (contemporary music, film and disk industries, electronic music industry). Like in speech synthesis, two major types of techniques are used for signal generation: descriptive-acoustic methods (rule-based formant synthesis) and signal processing methods (modification/concatenation of pre-recorded singing voices).

5.2.5 Future Directions

Prosodic modeling is probably the domain from which most of the improvements will come. In the long run it may be argued that the main problems to be solved deal mainly with mastering the linguistic and extra-linguistic phenomena related to prosody, which reflect problems of another kind, related to oral person-to-person and person-to-machine interactions.

Concerning the phonetic-acoustic generation process it may be foreseen that in the short run concatenative and articulatory syntheses will be boosted by the development of the microcomputer industry. By using off-the-shelf components it is already possible to implement a system using a large number of speech segments, with several variants that take into account contextual and prosodic effects, even for several speakers. This tendency can only be reinforced by the apparently unlimited evolution of computer speed and memory capacity, as well as by the fact that the computer industry not only provides the tools but also the market: speech synthesis nowadays must be considered to be as one of the most attractive aspects of virtual reality; it will benefit from the development of of multimedia and information highways.

5.3 Text Interpretation for TtS Synthesis

Richard Sproat

AT&T Bell Labs, Murray Hill, New Jersey, USA

The problem of converting text into speech for some language can naturally be broken down into two subproblems. One subproblem involves the conversion of linguistic parameter specifications (e.g., phoneme sequences, accentual parameters) into parameters (e.g., formant parameters, concatenative unit indices, pitch time/value pairs) that can drive the actual synthesis of speech. The other subproblem involves the computation of these linguistic parameter specifications from input text, which for the present discussion we will assume to be written in the standard orthographic representation for the language in question, and electronically coded in a standard scheme such as ASCII, ISO, JIS, BIG5, GB, and the like, depending upon the language. It is this second problem that is the topic of this section.

In any language, orthography is an imperfect representation of the underlying linguistic form. To illustrate this point, and to introduce some of the issues that we will discuss in this section, consider an English sentence such as *Give me a ticket to Dallas or give me back my money*: see Figure 5.2.

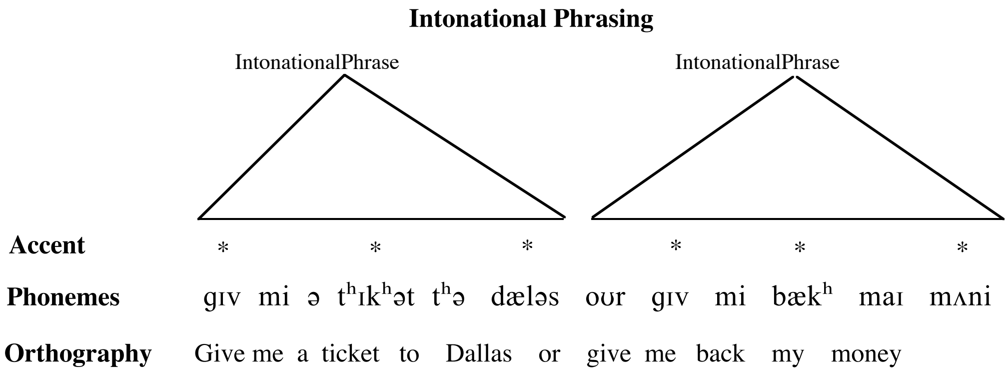

Figure 5.2: Some linguistic structures associated with the analysis of the sentence, "Give me a ticket to Dallas or give me back my money."

One of the first things that an English TtS system would need to do is tokenize the input into words: for English this is not generally difficult though for some other languages it is more complicated. A pronunciation then needs to be computed for each word; in English, given the irregularity of the orthography, this process involves a fair amount of lexical lookup though other processes are involved too. Some of the words in the sentence should be accented; in this particular case, a reasonable accentuation would involve accenting *content* words like *give*, *ticket*, *Dallas*, *back* and *money*, and leaving the other words unaccented. Then we might consider breaking the input into prosodic phrases: in this case, it would be reasonable to intone the sentence as if there were a comma between *Dallas* and *or*. Thus, various kinds of linguistic information need to be extracted from the text, but only in the case of word boundaries can this linguistic information be said to be represented directly in the orthography. In this survey I will focus on the topics of tokenization into words; the pronunciation of those words; the assignment of phrasal accentuation; and the assignment of prosodic phrases. An important area about which I will say little is what is often termed *text normalization*, comprising things like end-of-sentence detection, the expansion of abbreviations, and the treatment of acronyms and numbers.

5.3.1 Tokenization

As noted above, one of the first stages of analysis of the text input is the tokenization of the input into words. For many languages, including English, this problem is fairly easy in that one can to a first approximation assume that word boundaries coincide with whitespace or punctuation in the input text. In contrast, in many Asian languages the situation is not so simple, since spaces are never used in the orthographies of those languages to delimit words. In Chinese for example, whitespace generally only occurs in running text at paragraph boundaries. The Chinese alphabet consists of several thousand distinct elements, usually termed *characters*. With few exceptions, characters are monosyllabic. More controversially, one can also claim that most characters represent morphemes.

Just as words in English may consist of one or more morphemes so Chinese words may also consist of one or more morphemes. In a TtS system there are various reasons why it is important to segment Chinese text into words (as opposed to having the system read the input character-by-character). Probably the easiest of these to understand is that quite a few characters have more than one possible pronunciation, where the pronunciation chosen depends in many cases upon the particular word in which the character finds itself. A minimal requirement for word segmentation would appear to be an on-line dictionary that enumerates the word forms of the language. Indeed, virtually all Chinese segmenters reported in the literature contain a reasonably large dictionary (Chen & Liu, 1992; Wu & Tseng, 1993; Lin, Chiang, et al., 1993; Sproat, Shih, et al., 1994). Given a dictionary, however, one is still faced with the problem of how to use the lexical information to segment an input sentence: it is often the case that a sentence has more than one possible segmentation, so some method has to be employed to decide on the best analysis. Both heuristic (e.g., a greedy algorithm that finds the longest word at any point) and statistical approaches (algorithms that find the most probable sequence of words according to some model) have been applied to this problem.

While a dictionary is certainly a necessity for doing Chinese segmentation, it is not sufficient since in Chinese, as in English, any given text is likely to contain some words that are not found in the dictionary. Among these are words that are derived via morphologically productive processes, personal names and foreign names in transliteration. For morphologically complex forms, standard techniques for morphological analysis can be applied (Koskenniemi, 1983; Tzoukermann & Liberman, 1990; Karttunen, Kaplan, et al., 1992; Sproat, 1992), though some augmentation of these techniques is necessary in the case of statistical methods (Sproat, Shih, et al., 1994). Various statistical and non-statistical methods for handling personal and foreign names have been reported; see, for example, Chang, Chen, et al. (1992); Wang, Li, et al. (1992); Sproat, Shih, et al. (1994).

The period since the late 1980s has seen an explosion of work on the various problems of Chinese word segmentation, due in large measure to the increasing availability of large electronic corpora of Chinese text. Still, there is much work left to be done in this area, both in improving algorithms, and in the development of replicable evaluation criteria, the current lack of which makes fair comparisons of different approaches well-nigh impossible.

5.3.2 Word Pronunciation

Once the input is tokenized into words, the next obvious thing that must be done is to compute a pronunciation (or a set of possible pronunciations) for the words, given the orthographic representation of those words. The simplest approach is to have a set of *letter-to-sound rules* that simply map sequences of graphemes into sequences of phonemes, along with possible diacritic information, such as stress placement. This approach is naturally best suited to languages like Spanish or Finnish where there is a relatively simple relation between orthography and phonology. For languages like English, however, it has generally been recognized that a highly accurate word pronunciation module must contain a pronouncing dictionary that at the very least records words whose pronunciation could not be predicted on the basis of general rules.[1] Of course, the same problems of coverage as were noted in the Chinese segmentation problem also apply in the case of pronouncing dictionaries: many text words occur that are not to be found in the dictionary, the most important of these being morphological derivatives from known words, or previously unseen personal names.

[1] Some connectionist approaches to letter-to-sound conversion have attempted to replace traditional letter-to-sound rules with connectionist networks, and at the same time eschew the use of online dictionaries (for example, Sejnowski & Rosenberg, 1987). For English at least, these approaches would appear to have met with only limited success, however.

For morphological derivatives, standard techniques for morphological analysis can be applied to achieve a morphological decomposition for a word; see Allen, Hunnicutt, et al. (1987). The pronunciation of the whole can then in general be computed from the (presumably known) pronunciation of the morphological parts, applying appropriate phonological rules of the language. Morphological analysis is of some use in the prediction of name pronunciation too, since some names are derived from others via fairly productive morphological processes (cf., *Robertson* and *Robert*). However, this is not always the case, and one must also rely on other methods. One such method involves computing the pronunciation of a new name by analogy with the pronunciation of a similar name (Coker, Church, et al., 1990; Golding, 1991) (and see also Dedina & Nusbaum, 1991 for a more general application of analogical reasoning to word pronunciation). For example, if we have the name *Califano* in our dictionary and know its pronunciation, then we can compute the pronunciation of a hypothetical name *Balifano* by noting that both names share the final substring *alifano*: *Balifano* can then be pronounced on analogy by removing the phoneme /k/, corresponding to the letter *C* in *Califano*, and replacing it with the phoneme /b/. Yet another approach to handling proper names involves computing the language of origin of a name, typically by means of n-gram models of letter sequences for the various languages; once the origin of the name is guessed, *language*-specific pronunciation rules can be invoked to pronounce the name (Church, 1985; Vitale, 1991).

In many languages there are word forms that are inherently ambiguous in pronunciation, and for which a word pronunciation module as just described can only return a set of possible pronunciations, from which the most reasonable one must then be chosen. For example, the word *bass* rhymes with *lass* if it denotes a type of fish, and is homophonous with *base* if it denotes a musical range. An approach to this problem is discussed in Yarowsky (1994) (and see also Sproat, Hirschberg, et al., 1992). The method starts with a training corpus containing tagged examples in context of each pronunciation of a homograph. Significant local evidence (e.g., n-grams containing the homograph in question that are strongly associated to one or another pronunciation) and wide-context evidence (i.e., words that occur anywhere in the same sentence that are strongly associated to one of the pronunciations) are collected into a decision list, wherein each piece of evidence is ordered according to its strength (log likelihood of each pronunciation given the evidence). A novel instance of the homograph is then disambiguated by finding the strongest piece of evidence in the context in which the novel instance occurs, and letting that piece of evidence decide the matter. It is clear that the above-described method can also be applied to other formally similar problems in TtS, such as abbreviation expansion: for example is *St.* to be expanded as *Saint* or *Street*?

5.3.3 Accentuation

In many languages various words in a sentence are associated with *accents*, which are often manifested as upward or downward movements of fundamental frequency. Usually, not every word in the sentence bears an accent, however, and the decision of which words should be accented and which ones should not is one of the problems that must be addressed by a TtS system. More precisely, we will want to distinguish three levels of *prominence*, two being *accented* and *unaccented*, as just described, and the third being *cliticized*. Cliticized words are unaccented but additionally lack word stress, with the consequence that they tend to be durationally short.

A good first step in assigning accents is to make the accentual determination on the basis of broad lexical categories or parts of speech of words. Content words—nouns, verbs, adjectives and perhaps adverbs, tend in general to be accented; function words, including auxiliary verbs and prepositions tend to be deaccented; short function words tend to be cliticized. Naturally this presumes some method for assigning parts of speech, and in particular for disambiguating words like *can* which can be either content words (in this case, a verb or a noun), or function words (in this case, an auxiliary); fortunately, somewhat robust methods for part-of-speech tagging exist (e.g., Church, 1988). Of course, a finer-grained part-of-speech classification also reveals a finer-grained structure to the accenting problem. For example, the distinction between prepositions (*up the spout*) and particles (*give up*) is important in English since prepositions are typically deaccented or cliticized while particles are typically accented (Hirschberg, 1993).

But accenting has a wider function than merely communicating lexical category distinctions between words. In English, one important set of constructions where accenting is more complicated than what might be inferred from the above discussion are complex noun phrases—basically, a noun preceded by one or more adjectival or nominal modifiers. In a *discourse-neutral* context, some constructions are accented on the final word (*Madison* **Avenue**), some on the penultimate (**Wall** *Street*, *kitchen* **towel** *rack*), and some on an even earlier word (**sump** *pump factory*). Accenting on nominals longer than two words, is generally predictable given that one can compute the nominal's structure (itself a non-trivial problem), and given that one knows the accentuation pattern of the binary nominals embedded in the larger construction (Liberman & Prince, 1977; Liberman & Sproat, 1992; Sproat, 1994). Most linguistic work on nominal accent (e.g., Fudge, 1984; Liberman & Sproat, 1992, though see Ladd, 1984) has concluded that the primary determinants of accenting are semantic, but that within each semantic class there are lexically or semantically determined exceptions. For instance, righthand accent is often found in cases where the lefthand element denotes a location or time for the second element (cf. *morning* **paper**), but there are numerous lexical exceptions (**morning** *sickness*). Recent computational models—e.g., Monaghan

(1990); Sproat (1994)—have been partly successful at modeling the semantic and lexical generalizations; for example Sproat (1994) uses a combination of hand-built lexical and semantic rules, as well as a statistical model based on a corpus of nominals hand-tagged with accenting information.

Accenting is not only sensitive to syntactic structure and semantics, but also to properties of the discourse. One straightforward effect is *givenness*. In a case like *my son badly wants a dog, but I am* **allergic** *to dogs* where the second occurrence of *dogs* would often be deaccented because of the previous mention of *dog*. (See Hirschberg (1993) for a discussion of how to model this and other discourse effects, as well as the syntactic and semantic effects previously mentioned, in a working TtS module.) While humanlike accenting capabilities are possible in many cases, there are still many unsolved problems, a point we return to in the concluding subsection.

5.3.4 Prosodic Phrasing

The final topic that we address is the problem of chunking a long sentence into prosodic phrases. In reading a long sentence, speakers will normally break the sentence up into several phrases, each of which can be said to *stand alone* as an intonational unit. If punctuation is used liberally so that there are relatively few words between the commas, semicolons or periods, then a reasonable guess at an appropriate phrasing would be simply to break the sentence at the punctuation marks—though this is not always appropriate (O'Shaughnessy, 1989). The real problem comes when long stretches occur without punctuation; in such cases, human readers would normally break the string of words into phrases, and the problem then arises of where to place these breaks.

The simplest approach is to have a list of words, typically function words, that are likely indicators of good places to break (Klatt, 1987). One has to use some caution however, since while a particular function word like *and* may coincide with a plausible phrase break in some cases, in other cases it might coincide with a particularly *poor* place to break: *I was forced to sit through a dog and pony show that lasted most of Wednesday afternoon.*

An obvious improvement would be to incorporate an accurate syntactic parser and then derive the prosodic phrasing from the syntactic groupings: prosodic phrases usually do not coincide exactly with major syntactic phrases, but the two are typically not totally unrelated either. Prosodic phrasers that incorporate syntactic parsers are discussed in O'Shaughnessy (1989); Bachenko and Fitzpatrick (1990). O'Shaughnessy's system relies on a small lexicon of (mostly function) words that are reliable indicators of the beginnings of syntactic groups: articles such as *a* or *the* clearly indicate the beginnings of noun groups, for example. This lexicon is augmented by suffix-stripping rules that allow for part-of-speech assignment to words where this information can be predicted from the morphology. A bottom-up parser is then used to construct phrases based upon the syntactic-group-indicating words. Bachenko and Fitzpatrick employ a somewhat more sophisticated deterministic syntactic parser (FIDDITCH Hindle, 1983) to construct a syntactic analysis for a sentence; the syntactic phrases are then transduced into prosodic phrases using a set of heuristics.

But syntactic parsing *sensu stricto* may not be necessary in order to achieve reasonable predictions of prosodic phrase boundaries. Wang and Hirschberg (1992) report on a corpus-based statistical approach that uses CART (Breiman, Friedman, et al., 1984; Riley, 1989) to train a decision tree on transcribed speech data. In training, the dependent variable was the human prosodic phrase boundary decision, and the independent variables were generally properties that were computable automatically from the text including: part of speech sequence around the boundary; the location of the edges of long noun phrases (as computable from automatic methods such as Church, 1988; Sproat, 1994); distance of the boundary from the edges of the sentence, and so forth.

5.3.5 Future Directions

This section has given an overview of a selected set of the problems that arise in the conversion of textual input into a linguistic representation suitable for input to a speech synthesizer, and has outlined a few solutions to these problems. As a result of these solutions, current *high-end* TtS systems produce speech output that is quite intelligible and in many cases quite natural. For example, in English it is possible to produce TtS output where the vast majority of words in a text are correctly pronounced, where words are mostly accented in a plausible fashion, and where prosodic phrase boundaries are chosen at mostly reasonable places. Nonetheless, even the best systems make mistakes on unrestricted text, and there is much room for improvement in the approaches taken to solving the various problems, though one can of course often improve performance marginally by tweaking existing approaches.

Perhaps the single most important unsolved issue that affects performance on many of the problems discussed in this section is that full machine *understanding* of unrestricted text is currently not possible, and so TtS systems can fairly be said to not know what they are talking about. This point comes up rather clearly in the treatment of accenting in English, though the point could equally well be made in other areas. As we noted above, previously mentioned items are often deaccented, and this would be appropriate for the second occurrence of *dog* in the sentence *my son badly wants a dog, but I am* **allergic** *to dogs*. But a moment's reflection will reveal that what is crucial is not the repetition of the word *dog*, but rather the repetition of the concept *dog*. That what is relevant is semantic or conceptual categories and not simply words becomes clear when one considers that one also would often deaccent a word if a conceptual supercategory of that word had been previously mentioned: *My son wants a labrador, but I'm* **allergic** *to dogs*. Various solutions involving semantic networks (such as WordNet) might be contemplated, but so far no promising results have been reported.

Note that *message*-to-speech systems have an advantage over *text*-to-speech systems precisely in that message-to-speech systems in some sense *know* what they are talking about since one can code as much semantic knowledge into the initial message as one desires. But TtS systems must compute everything from orthography which, as we have seen, is not very informative about a large number of linguistic properties of speech.

5.4 Spoken Language Generation

Kathleen R. McKeown[a] & Johanna D. Moore[b]

[a] Columbia University, New York, New York, USA
[b] University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Interactive natural language capabilities are needed for a wide range of today's intelligent systems: expert systems must explain their results and reasoning, intelligent assistants must collaborate with users to perform tasks, tutoring systems must teach domain concepts and critique students' problem-solving strategies, and information delivery systems must help users find and make sense of the information they need. These applications require that a system be capable of generating coherent multisentential responses, and interpreting and responding to users' subsequent utterances in the context of the ongoing interaction.

Spoken language generation allows for provision of responses as part of an interactive human-machine dialogue, where speech is one medium for the response. This research topic draws from the fields of both natural language generation and speech synthesis. It differs from synthesis in that speech is generated from an abstract representation of concepts rather than from text. While a relatively under-emphasized research problem, the ability to generate spoken responses is clearly crucial for interactive situations, in particular when:

1. the user's hands and/or eyes are busy;
2. screen real estate is at a premium;
3. time is critical; or
4. system and user are communicating via a primarily audio channel such as the telephone.

Like written language generation, spoken language generation requires determining what concepts to include and how to realize them in words, but critically also requires determining intonational form. Several problems are particularly pertinent to the spoken context:

- The need to model and use knowledge about hearer goals, hearer background, and past discourse in determining content and form of a response. While the written context can include a general audience (e.g., for report generation), responses in an interactive dialog are intended for a particular person and to be useful, must take that person into account.

- What kind of language should be generated given a spoken context? Given the lack of visual memory that a text provides, the form required for speech is likely to be quite different from that found in text.
- In the processing of determining the content and form of the response, how can a system provide information to control intonation, which is known to provide crucial clues as to intended meaning.

5.4.1 State of the Art

The field of spoken language generation is in its infancy, with very few researchers working on systems that deal with all aspects of producing spoken language responses, i.e., determining what to say, how to say it, and how to pronounce it. In fact, in spoken language systems, such as the ARPA Air Travel Information Service (ATIS), the focus has been on correctly interpreting the spoken request, relying on direct display of database search results and minimal response generation capabilities. However, much work on written response generation as part of interactive systems is directly applicable to spoken language generation; the same problems must be addressed in an interactive spoken dialog system. Within speech synthesis, research on controlling intonation to signal meaning and discourse structure is relevant to the problem. This work has resulted in several concept to speech systems.

Interactive Systems

Research in natural language understanding has shown that coherent discourse has structure, and that recognizing the structure is a crucial component of comprehending the discourse (Grosz & Sidner, 1986; Hobbs, 1993; Moore & Pollack, 1992). Thus, generation systems participating in dialog must be able to select and organize content as part of a larger discourse structure and convey this structure, as well as the content, to users. This has led to the development of several plan-based models of discourse, and to implemented systems that are capable of participating in a written interactive dialogue with users (Cawsey, 1993; Maybury, 1992; Moore, 1995).

Two aspects of discourse structure are especially important for spoken language generation. First is *intentional structure*, which describes the roles that discourse actions play in the speaker's communicative plan to achieve desired effects on the hearer's mental state. Moore and Paris (1993) have shown that intentional structure is crucial for responding effectively to questions that address a previous utterance: without a record of what an utterance was intended to achieve, it is impossible to elaborate or clarify that utterance. In addition, information about speaker intentions has been shown to be an important factor in selecting appropriate lexical items, including discourse cues (e.g., *because*, *when*, *although*; Moser & Moore, 1995a; Moser & Moore, 1995b) and scalar terms (e.g., *difficult*, *easy*; Elhadad, 1992).

Second is *attentional structure* (Carberry, 1983; Grosz, 1977; Grosz & Sidner, 1986; Gordon, Grosz, et al., 1993; Sidner, 1979), which contains information about the objects, properties, relations, and discourse intentions that are most salient at any given point in the discourse. In natural discourse, humans *focus* or *center* their attention on a small set of entities and attention shifts to new entities in predictable ways. Many generation systems track focus; of attention as the discourse as a whole progresses as well as during the construction of its individual responses (McCoy & Cheng, 1990; McKeown, 1985; Sibun, 1992). Focus has been used to determine when to pronimalize, to make choices in syntactic form (e.g., active vs. passive), and to appropriately mark changes in topic, e.g., the introduction of a new topic or return to a previous topic (Cawsey, 1993). Once tracked, such information would be available for use in speech synthesis as described below.

Another important factor for response generation in interactive systems is the ability to tailor responses based on a model of the intended hearer. Researchers have developed systems capable of tailoring their responses to the user's background (Cohen, Jones, et al., 1989), level of expertise (Paris, 1988), goals (McKeown, 1988), preferences (Chu-Carroll & Carberry, 1994), or misconceptions (McCoy, 1986). In addition, generating responses that the user will understand requires that the system use terminology that is familiar to the user (McKeown, Robin, et al., 1993).

Controlling Intonation to Signal Meaning in Speech Generation

Many studies have shown that intonational information is crucial for conveying intended meaning in spoken language (Butterworth, 1975; Hirschberg & Pierrehumbert, 1986; Silverman, 1987). For example, Pierrehumbert and Hirschberg (1990) identify how pitch accents indicate the information status of an item (e.g., given/new) in discourse, how variations in intermediate phrasing can convey structural relations among elements of a phrase, and how variation in pitch range can indicate topic changes. In later work, Hirschberg and Litman (1993) show that pitch accent and prosodic phrasing distinguish between discourse and sentential uses of cue phrases (e.g., *now* and *well*), providing a model for selecting appropriate intonational features when generating these cue phrases in synthetic speech. There have been only a few interactive spoken language systems that exploit intonation to convey meaning. Those that do generate speech from an abstract representation of content that allows tracking focus, given/new information, topic switches, and discourse segmentation (for one exception, see the Telephone Enquiry System (TES) (Witten & Madams, 1977) where text was augmented by hand to include a coded intonation scheme). The Speech Synthesis from Concept (SSC) system, developed by Young and Fallside (1979) showed how syntactic structure could be used to aid in decisions about accenting and phrasing. Davis and Hirschberg (1988) developed a *message-to-speech* system that uses structural, semantic, and discourse information to control assignment of pitch range, accent placement, phrasing and pause. The result is a system that generates spoken directions with appropriate intonational features given start and end coordinates on a map. The generation of contrastive intonation is being explored in a medical information system, where full answers to yes-no questions are generated (Prevost & Steedman, 1994; Prevost, 1995). It is only in this last system that language generation techniques (e.g., a generation grammar) are fully explored. Other recent approaches to concept to speech generation can also be found (Horne & Filipsson, 1994; House & Youd, 1990).

5.4.2 Future Directions

Spoken language generation is a field in which more remains to be done than has been done to date. Although response generation is a critical component of interactive spoken language systems, and of any human computer interface, many current systems assume that once a spoken utterance is interpreted, the response can be made using the underlying system application (e.g., the results of a database search) and commercial speech synthesizers. If we are to produce effective spoken language human computer interfaces, then a concerted effort on spoken language generation must be pursued. Such interfaces would be clearly useful in applications such as task assisted instruction giving (e.g., equipment repair), telephone information services, medical information services (e.g., updates during surgery), commentary on animated information (e.g., animated algorithms), spoken translation, or summarization of phone transcripts.

Interaction Between Generation and Synthesis

To date, research on the interaction between discourse features and intonation has been carried out primarily by speech synthesis groups. While language generation systems often track the required discourse features, there have been few attempts to integrate language generation and speech synthesis. This would require the generation system to provide synthesis with the parameters needed to control intonation. By providing more information than is available to a TtS synthesis system and by requiring language generation to refine representations of discourse features for intonation, research in both fields will advance.

Generating Language Appropriate to Spoken Situations

Selecting the words and syntactic structure of a generated response has been explored primarily from the point of view of written language (see Hovy, this volume). If a response is to be spoken, however, it will have different characteristics than does written language. For example, it is unlikely that long complex sentences will be appropriate without the visual, written context. Research is needed that incorporates the results of work in psycholinguistics on constraints on spoken language form (Levelt, 1989) into generation systems, that identifies further constraints on variability in surface form, and that develops both grammars and lexical choosers that produce the form of language required in a spoken context. While there has been some work on the development of incremental, real-time processes for generation of spoken language (De Smedt, 1990; McDonald, 1983), more work is needed on constraints.

Influence of Discourse History

When generation takes place as part of an interactive dialogue system, responses must be sensitive to what has already been said in the current session and to the individual user. This influences the content of the response; the system should relate new information to recently conveyed material and avoid repeating old material that would distract the user from what is new. The discourse history also influences the form of the response; the system must select vocabulary that the user can understand. Furthermore, knowledge about what information is new, or not previously mentioned, and what information is given, or available from previous discourse, influences the use of anaphoric expressions as well as word ordering. There has been some work on generating referring expressions appropriate to context, e.g., pronouns and definite descriptions (McDonald, 1980, pp. 218–220; Dale, 1989; Granville, 1984). In addition, there has been some work on producing responses to follow-up questions (Moore & Paris, 1993), on generating alternative explanations when a first attempt is not understood (Moore, 1989), and on issues related to managing the initiative in a dialogue (Haller, 1994; McRoy, 1995). However, much remains to be done, particularly in dialogs involving collaborative problem solving or in cases where the dialog involves mixed initiative.

Coordination with Other Media

When response generation is part of a larger interactive setting, including speech, graphics, animation, as well as written language, a generator must coordinate its tasks with other components. For example, which information in the selected content should appear in language and which in graphics? If speech and animation are used, how are they to be coordinated temporally (e.g., how much can be said during a given scene)? What parameters used during response generation tasks should be made available to a speech component? These are issues that have only recently surfaced in the research community.

Evaluating Spoken Language Generation

There has been very little work on how to measure when a generation system is successful. Possibilities include evaluating how well a user can complete a task which requires interaction with a system that generates responses, asking users to indicate satisfaction with system responses, performing a preference analysis between different types of text, degrading a response generation system and testing user satisfaction, and evaluating system generation against a target case, among others. Each one of these has potential problems. For example, task completion measures definitely interact with the front end interface: that is, how easy is it for a user to request the information needed? Thus, it would be helpful to have interaction between computer scientists that build the systems and psychologists, who are better trained in creating valid evaluation techniques to produce better ways for understanding how well a generation system works.

5.5 Chapter References

Abe, M., Nakamura, S., Shikano, K., and Kuwabara, H. (1990). Voice conversion through vector quantization. *Journal of the Acoustical Society of Japan*, E-11:71–76.

ACL (1986). *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, Columbia University, New York. Association for Computational Linguistics.

Allen, J., Hunnicutt, M. S., and Klatt, D. (1987). *From text to speech—the MITalk system.* MIT Press, Cambridge, Massachusetts.

Atal, B. S. and Remde, J. R. (1982). A new model of LPC excitation for producing natural-sounding speech at low bit rates. In *Proceedings of the 1982 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 614–617. Institute of Electrical and Electronic Engineers.

Bachenko, J. and Fitzpatrick, E. (1990). A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, 16:155–170.

Bailly, G. and Benoît, C., editors (1990). *Proceedings of the First ESCA Workshop on Speech Synthesis*, Autrans, France. European Speech Communication Association.

Bailly, G. and Benoît, C., editors (1992). *Talking Machines: Theories, Models, and Designs.* Elsevier Science.

Bartkova, K. and Sorin, C. (1987). A model of segmental duration for speech synthesis in French. *Speech Communication*, 6:245–260.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth & Brooks, Pacific Grove, California.

Butterworth, B. (1975). Hesitation and semantic planning in speech. *Journal of Psycholinguistic Research*, 4:75–87.

Campbell, W. N. (1992). Syllable-based segmental duration. In Bailly, G. and Benoît, C., editors, *Talking Machines: Theories, Models, and Designs*, pages 211–224. Elsevier Science.

Campbell, W. N. and Isard, S. D. (1991). Segment durations in a syllable frame. *Journal of Phonetics Computation Speech and Language*, 19:37–47.

Carberry, S. (1983). Tracking user goals in an information-seeking environment. In *Proceedings of the Third National Conference on Artificial Intelligence*, pages 59–63, Washington, DC.

Carlson, R. and Granström, B. (1986). A search for durational rules in a real-speech data base. *Phonetica*, 43:140–154.

Cawsey, A. (1993). *Explanation and Interaction: The Computer Generation of Explanatory Dialogues*. MIT Press.

Chang, J.-S., Chen, S.-D., Zheng, Y., Liu, X.-Z., and Ke, S.-J. (1992). Large-corpus-based methods for Chinese personal name recognition. *Journal of Chinese Information Processing*, 6(3):7–15.

Chen, K.-J. and Liu, S.-H. (1992). Word identification for Mandarin Chinese sentences. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 101–107, Nantes, France. ACL.

Chu-Carroll, J. and Carberry, S. (1994). A plan-based model for response generation in collaborative task-oriented dialogues. In *Proceedings of the National Conference on Artificial Intelligence*, pages 799–805, Menlo Park, California. AAAI Press.

Church, K. (1985). Stress assignment in letter to sound rules for speech synthesis. In *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics*, pages 246–253, University of Chicago. Association for Computational Linguistics.

Church, K. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136–143, Austin, Texas. ACL.

Cohen, R., Jones, M., Sanmugasunderam, A., Spencer, B., and Dent, L. (1989). Providing responses specific to a user's goals and background. *International Journal of Expert Systems*, 2(2):135–162.

Coker, C., Church, K., and Liberman, M. (1990). Morphology and rhyming: Two powerful alternatives to letter-to-sound rules for speech synthesis. In Bailly, G. and Benoît, C., editors, *Proceedings of the First ESCA Workshop on Speech Synthesis*, pages 83–86, Autrans, France. European Speech Communication Association.

COLING (1992). *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France. ACL.

Dale, R. (1989). Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 68–75, Vancouver, British Columbia. Association for Computational Linguistics.

d'Alessandro, C. and Mertens, P. (1995). Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language*, 9:257–288.

Davis, J. R. and Hirschberg, J. (1988). Assigning intonational features in synthesized spoken directions. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 187–193, SUNY, Buffalo, New York. Association for Computational Linguistics.

De Smedt, K. J. M. J. (1990). IPF: an incremental parallel formulator. In Dale, R., Mellish, C. S., and Zock, M., editors, *Current Research in Natural Language Generation*. Academic Press, London.

Dedina, M. and Nusbaum, H. (1991). PRONOUNCE: a program for pronunciation by analogy. *Computer Speech and Language*, 5:55–64.

Dutoit, T. and Leich, H. (1993). MBR-PSOLA: Text-to-speech synthesis based on an MBEre-synthesis of the segments database. *Speech Communication*, 13:432–440.

Elhadad, M. (1992). *Using Argumentation to Control Lexical Choice: A Functional Unification-Based Approach*. PhD thesis, Computer Science Department, Columbia University.

ESCA (1994). *Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*, New Paltz, New York. European Speech Communication Association.

Fant, G., Liljencrants, J., and Lin, Q. (1985). A four parameter model of glottal flow. *Speech Transactions Laboratory Quarterly and Status Report*, 1985(4):1–13.

Flanagan, J. L. and Rabiner, L. R., editors (1973). *Speech Synthesis*. Dowden, Hutchinson & Ross.

Fudge, E. (1984). *English Word-Stress*. Allen and Unwin, London.

Fujisaki, H. (1992). Modeling the process of fundamental frequency contour generation. In *Speech perception, production and linguistic structure*, pages 313–326. Ohmsha IOS Press.

Fujisaki, T. and Kawai, H. (1988). Realization of linguistic information in the voice fundamental frequency contour of the spoken Japanese. In *Proceedings of the 1988 International Conference on Acoustics, Speech, and Signal Processing*, pages 663–666, New York.

Golding, A. (1991). *Pronouncing Names by a Combination of Case-Based and Rule-Based Reasoning*. PhD thesis, Stanford University.

Gordon, P. C., Grosz, B. J., and Gilliom, L. A. (1993). Prounouns, names and the centering of attention in discourse. *Cognitive Science*, 17(3):311–348.

Granström, B. and Nord, L. (1992). Neglected dimensions in speech synthesis. *Speech Communication*, 11:459–462.

Granville, R. (1984). Controlling lexical substitution in computer text generation. In *Proceedings of the 10th International Conference on Computational Linguistics*, pages 381–384, Stanford University, California. ACL.

Grosz, B. J. (1977). The representation and use of focus in dialogue understanding. Technical Report 151, SRI International, Menlo Park, California.

Grosz, B. J. and Sidner, C. L. (1986). Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Hakoda, K. and Sato, H. (1980). Prosodic rules in connected speech synthesis. *Trans. IECE*, pages 715–722.

Haller, S. M. (1994). Recognizing digressive questions using a model for interactive generation. In *Proceedings of the 7th International Workshop on Natural Language Generation*, pages 181–188, Kinnebunkport, Maine.

Hart, J., Collier, R., and Cohen, A., editors (1990). *A perceptual study of intonation*. Cambridge University Press, Cambridge, England.

Hindle, D. (1983). User manual for Fidditch, a deterministic parser. Technical Report Technical Memorandum 7590-142, Naval Research Laboratory.

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

Hirschberg, J. and Litman, D. (1993). Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–530.

Hirschberg, J. and Pierrehumbert, J. (1986). The intonational structuring of discourse. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pages 136–144, Columbia University, New York. Association for Computational Linguistics.

Hobbs, J. R. (1993). Intention, information, and structure in discourse. In *Proceedings of the NATO Advanced Research Workshop on Burning Issues in Discourse*, pages 41–66, Maratea, Italy.

Horne, M. and Filipsson, M. (1994). Computational extraction of lexico-grammatical information for generation of Swedish intonation. In *Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*, pages 220–223, New Paltz, New York. European Speech Communication Association.

House, J. and Youd, N. (1990). Contextually appropriate intonation in speech synthesis. In Bailly, G. and Benoît, C., editors, *Proceedings of the First ESCA Workshop on Speech Synthesis*, pages 185–188, Autrans, France. European Speech Communication Association.

ICSLP (1992). *Proceedings of the 1992 International Conference on Spoken Language Processing*, Banff, Alberta, Canada. University of Alberta.

Iwahashi, N. and Sagisaka, Y. (1993). Duration modeling with multiple split regression. In *Eurospeech '93, Proceedings of the Third European Conference on Speech Communication and Technology*, volume 1, pages 329–332, Berlin. European Speech Communication Association.

Kaiki, N., Takeda, K., and Sagisaka, Y. (1992). Linguistic properties in the control of segmental duration for speech synthesis. In Bailly, G. and Benoît, C., editors, *Talking Machines: Theories, Models, and Designs*, pages 255–264. Elsevier Science.

Karttunen, L., Kaplan, R. M., and Zaenen, A. (1992). Two-level morphology with composition. In *Proceedings of the 14th International Conference on Computational Linguistics*, volume 1, pages 141–148, Nantes, France. ACL.

Kato, H., Tsuzaki, M., and Sagisaka, Y. (1992). Acceptability and discrimination threshold for distortion of duration in Japanese words. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, volume 1, pages 507–510, Banff, Alberta, Canada. University of Alberta.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67:971–995.

Klatt, D. H. (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, 82(3):737–793.

Klatt, D. H. and Klatt, L. C. (1990). Analysis, synthesis a, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87:820–857.

Koskenniemi, K. (1983). *Two-Level Morphology: a General Computational Model for Word-Form Recognition and Production.* PhD thesis, University of Helsinki.

Publications of the Department of General Linguistics,University of Helsinki, No. 11. Helsinki.

Kraft, V. and Portele, T. (1995). Quality evaluation of five German speech synthesis systems. *Acta Acustica*, 3:351–365.

Ladd, D. R. (1984). English compound stress. In Gibbon, D. and Richter, H., editors, *Intonation, Accent and Rhythm*, pages 253–266. W. de Gruyter, Berlin.

Laroche, J., Stylianou, Y., and Moulines, E. (1993). HNS: Speech modification based on a harmonic + noise model. In *Proceedings of the 1993 International Conference on Acoustics, Speech, and Signal Processing*, pages 550–553.

Levelt, W. (1989). *Speaking: from intention to articulation.* MIT Press, Cambridge, Massachusetts.

Liberman, M. and Pierrehumbert, J. B. (1984). Intonational invariance under changes in pitch range and length. In *Language Sound Structure*, pages 157–233. MIT Press.

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

Liberman, M. and Sproat, R. (1992). The stress and structure of modified noun phrases in English. In Szabolcsi, A. and Sag, I., editors, *Lexical Matters*. CSLI (University of Chicago Press).

Lin, M.-Y., Chiang, T.-H., and Su, K.-Y. (1993). A preliminary study on unknown word problem in Chinese word segmentation. In *ROCLING 6*, pages 119–141. ROCLING.

Maeda, S. (1976). *A characterization of American English intonation.* PhD thesis, MIT.

Matsumoto, H., Maruyama, Y., and Inoue, H. (1994). Voice quality conversion based on supervised spectral mapping. *Journal of the Acoustical Society of Japan*, E. In press.

Maybury, M. T. (1992). Communicative acts for explanation generation. *International Journal of Man-Machine Studies*, 37(2):135–172.

McCoy, K. F. (1986). The ROMPER system: Responding to object-related misconceptions using perspective. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, Columbia University, New York. Association for Computational Linguistics.

McCoy, K. F. and Cheng, J. (1990). Focus of attention: Constraining what can be said next. In Paris, C. L., Swartout, W. R., and Mann, W. C., editors, *Natural Language Generation in Artificial Intelligence and Computational Linguistics*, pages 103–124. Kluwer Academic, Boston.

McDonald, D. D. (1980). *Natural Language Production as a Process of Decision Making Under Constraint*. PhD thesis, Department of Computer Science and Electrical Engineering, Massachusetts Institute of Technology.

McDonald, D. D. (1983). Description directed control: its implications for natural language generation. In Grosz, B. J., Sparck Jones, K., and Webber, B. L., editors, *Readings in Natural Language Processing*. Morgan Kaufmann Publishers, Inc.

McKeown, K. R. (1985). *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Studies in Natural Language Processing. Cambridge University Press.

McKeown, K. R. (1988). Generating goal-oriented explanations. *International Journal of Expert Systems*, 1(4):377–395.

McKeown, K. R., Robin, J., and Tanenblatt, M. (1993). Tailoring lexical choice to the user's vocabulary in multimedia explanation generation. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 226–234, Ohio State University. Association for Computational Linguistics.

McRoy, S. (1995). The repair of speech act misunderstandings by abductive inference. *Computational Linguistics*. In press.

Monaghan, A. (1990). Rhythm and stress in speech synthesis. *Computer Speech and Language*, 4:71–78.

Moore, J. D. (1989). Responding to "Huh?": Answering vaguely articulated follow-up questions. In *Proceedings of the Conference on Human Factors in Computing Systems*, pages 91–96, Austin, Texas.

Moore, J. D. (1995). *Participating in Explanatory Dialogues: Interpreting and Responding to Questions in Context*. MIT Press.

Moore, J. D. and Paris, C. L. (1993). Planning texts for advisory dialogues: Capturing intentional and rhetorical information. *Computational Linguistics*, 19(4):651–694.

Moore, J. D. and Pollack, M. E. (1992). A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics*, 18(4):537–544.

Moser, M. and Moore, J. D. (1995a). Investigating cue selection and placement in tutorial discourse. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, MIT. Association for Computational Linguistics.

Moser, M. and Moore, J. D. (1995b). Using discourse analysis and automatic text generation to study discourse cue usage. In *Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453–468.

Moulines, E. and Sagisaka, Y. (1995). Voice conversion: State of the art and perspectives. *Speech Communication*, 16(2). Guest editors.

Nakajima, S. and Hamada, H. (1988). Automatic generation of synthesis units based on context oriented clustering. In *Proceedings of the 1988 International Conference on Acoustics, Speech, and Signal Processing*, pages 659–662, New York. Institute of Electrical and Electronic Engineers.

Olive, J. P., Greenwood, A., and Coleman, J. (1993). *Acoustics of American English Speech, A Dynamic Approach*. Springer-Verlag.

O'Shaughnessy, D. (1989). Parsing with a small dictionary for applications such as text to speech. *Computational Linguistics*, 15:97–108.

Paris, C. L. (1988). Tailoring object descriptions to the user's level of expertise. *Computational Linguistics*, 14(3):64–78.

Parthasarathy, S. and Coker, C. H. (1992). Automatic estimation of articulatory parameters. *Computer Speech and Language*, 6:37–75.

Pierrehumbert, J. (1981). Synthesizing intonation. *Journal of the Acoustical Society of America*, 70:985–995.

Pierrehumbert, J. and Hirschberg, J. (1990). The meaning of intonational contours in interpretation of discourse. In Cohen, P. R., Morgan, J., and Pollack, M. E., editors, *Intentions in Communication*, pages 271–311. MIT Press, Cambridge, Massachusetts.

Pitrelli, J., Beckman, M., and Hirschberg, J. (1994). Evaluation of prosodic transcription labelling in the ToBI framework. In *Proceedings of the 1994 International Conference on Spoken Language Processing*, pages 123–126, Yokohama, Japan.

Prevost, S. A. (expected 1995). *Intonation, Context and Contrastiveness in Spoken Language Generation*. PhD thesis, University of Pennsylvania, Philadelphia, Pa.

Prevost, S. A. and Steedman, M. J. (1994). Specifying intonation from context for speech synthesis. *Speech Communication*, 15(1-2).

Richard, G. and d'Alessandro, C. (1994). Time-domain analysis-synthesis of the aperiodic component of speech signals. In *Proceedings of the ESCA Workshop on Speech Synthesis*, pages 5–8.

Riley, M. (1989). Some applications of tree-based modelling to speech and language. In *Proceedings of the Second DARPA Speech and Natural Language Workshop*, Cape Cod, Massachusetts. Defense Advanced Research Projects Agency.

Riley, M. D. (1992). Tree-based modeling of segmental durations. In Bailly, G. and Benoît, C., editors, *Talking Machines: Theories, Models, and Designs*, pages 265–273. Elsevier Science.

Sagisaka, Y., Kaiki, N., Iwahashi, N., and Mimura, K. (1992). ATR ν-Talk speech synthesis system. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, volume 1, pages 483–486, Banff, Alberta, Canada. University of Alberta.

Sejnowski, T. and Rosenberg, C. (1987). Parallel networks that learn to pronounce English text. *Complex Systems*, 1.

Sibun, P. (1992). Generating text without trees. *Computational Intelligence*, 8(1):102–122.

Sidner, C. L. (1979). *Toward a Computational Theory of Definite Anaphora Comprehension in English Discourse*. PhD thesis, Massachusetts Institute of Technology, Cambridge, Mass.

Silverman, K. (1987). *The structure and processing of fundamental frequency contours*. PhD thesis, Cambridge University, Cambridge, England.

Sproat, R. (1992). *Morphology and Computation*. MIT Press, Cambridge, Massachusetts.

Sproat, R. (1994). English noun-phrase accent prediction for text-to-speech. *Computer Speech and Language*, 8:79–94.

Sproat, R., Hirschberg, J., and Yarowsky, D. (1992). A corpus-based synthesizer. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, volume 1, pages 563–566, Banff, Alberta, Canada. University of Alberta.

Sproat, R., Shih, C., Gale, W., and Chang, N. (1994). A stochastic finite-state word-segmentation algorithm for Chinese. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 66–73, Las Cruces, New Mexico. Association for Computational Linguistics.

Stevens, K. N. and Bickley, C. (1991). Constraints among parameters simplify control of Klatt formant synthesizer. *Phonetics*, 19:161–174.

Sundberg, J. (1987). *The science of the singing voice*. Northern Illinois University Press, Dekalb, Illinois.

Takeda, K., Abe, K., and Sagisaka, Y. (1992). On the basic scheme and algorithms in non-uniform unit speech synthesis. In Bailly, G. and Benoît, C., editors, *Talking Machines: Theories, Models, and Designs*, pages 93–105. Elsevier Science.

Traber, C. (1992). F0 generation with a database of natural F0 pattern and with a neural network. In Bailly, G. and Benoît, C., editors, *Talking Machines: Theories, Models, and Designs*, pages 287–304. Elsevier Science.

Tzoukermann, E. and Liberman, M. Y. (1990). A finite-state morphological processor for Spanish. In Karlgren, H., editor, *Proceedings of the 13th International Conference on Computational Linguistics*, volume 3, pages 277–286, Helsinki. ACL.

Umeda, N. (1975). Vowel duration in American English. *Journal of the Acoustical Society of America*, 58(2):434–445.

Valbret, H., Moulines, E., and Tubach, J. (1992). Voice transformation using PSOLA. *Speech Communication*, 11:175–187.

Van Bezooijen, R. and Pols, L. (1990). Evaluation of text-to-speech systems: some methodological aspects. *Speech Communication*, 9:263–270.

Van Santen, J., Sproat, R., Olive, J., and Hirshberg, J., editors (1995). *Progress in Speech Synthesis*. Springer Verlag, New York.

Van Santen, J. P. H. (1992). Contextual effects on vowel duration. *Speech Communication*, 11:513–546.

Van Santen, J. P. H. (1993). Perceptual experiment for diagnostic testing of text-to-speech systems. *Computer Speech and Language*, 7:49–100.

Van Santen, J. P. H. (1994). Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, 8:95–128.

Vitale, T. (1991). An algorithm for high accuracy name pronunciation by parametric speech synthesizer. *Computational Linguistics*, 17:257–276.

Wang, L.-J., Li, W.-C., and Chang, C.-H. (1992). Recognizing unregistered names for Mandarin word identification. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 1239–1243, Nantes, France. ACL.

Wang, M. Q. and Hirschberg, J. (1992). Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, 6:175–196.

Witten, L. and Madams, P. (1977). The telephone inquiry service: a man-machine system using synthetic speech. *International Journal of Man-Machine Studies*, 9:449–464.

Wu, Z. and Tseng, G. (1993). Chinese text segmentation for text retrieval: Achievements and problems. *Journal of the American Society for Information Science*, 44(9):532–542.

Yamashita, Y., Tanaka, M., Amako, Y., Nomura, Y., Ohta, Y., Kitoh, A., Kakusho, O., and Mizoguchi, R. (1993). Tree-based approaches to automatic generation of speech synthesis rules for prosodic parameters. *Trans. IEICE*, E76-A(11):1934–1941.

Yarowsky, D. (1994). Homograph disambiguation in speech synthesis. In *Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*, pages 244–247, New Paltz, New York. European Speech Communication Association.

Young, S. J. and Fallside, F. (1979). Speech synthesis from concept: a method for speech output from information systems. *Journal of the Acoustic Society of America*, 66(3):685–695.