Issues in Text-to-Speech Conversion for Mandarin

Chilin Shih (石基琳) & Richard Sproat (史伯樂)
Speech Synthesis Research Department
Bell Laboratories
Lucent Technologies
700 Mountain Avenue, Room 2d-453/2d-451
Murray Hill, NJ, USA, 07974

{cls,rws}@bell-labs.com

Abstract

Research on text-to-speech (TTS) conversion for Mandarin Chinese is a much younger enterprise than comparable research for English or other European languages. Nonetheless, impressive progress has been made over the last couple of decades, and Mandarin Chinese systems now exist which approach, or in some ways even surpass in quality available systems for English. This article has two goals. The first is to summarize the published literature on Mandarin synthesis, with a view to clarifying the similarities or differences among the various efforts. One property shared by a great many systems is the dependence on the syllable as the basic unit of synthesis. We shall argue that this property stems both from the accidental fact that Mandarin has a small number of syllable types, and from traditional Sinological views of the linguistic structure of Chinese. Despite the popularity of the syllable, though, there are problems with using it as the basic synthesis unit, as we shall show. The second goal is to describe in more detail some specific problems in text-to-speech conversion for Mandarin, namely text analysis, concatenative unit selection, segmental duration and tone and intonation modeling. We illustrate these topics by describing our own work on Mandarin synthesis at Bell Laboratories. The paper starts with an introduction to some basic concepts in speech synthesis, which is intended as an aid to readers who are less familiar with this area of research.

Keywords: Chinese speech synthesis, text analysis, concatenative units, duration, tone and intonation.

PART ONE: A GENERAL OVERVIEW

1 Introduction

Text-to-speech synthesis — TTS — is the automatic conversion of a text into speech that resembles, as closely as possible, a native speaker of the language reading that text. In all current TTS systems, the text is in a digitally coded format, such as ASCII for English, or Big 5 or GuoBiao (GB) for Chinese.[1] We take it as axiomatic that a complete text-to-speech system for a language must be able to handle text as it is normally written in that language, using the standard orthography. Thus, a TTS system for Chinese that can only accept input in pinyin romanization is not a true TTS system. More generally, text contains input besides ordinary words. A typical Chinese newspaper article may contain numbers written in Arabic numerals, percentages and dollar amounts, non-Chinese words written in Roman orthography, and so on. A full text-to-speech system must be able to do something reasonable with these kinds of input also.

A closer examination of the problem of converting from text into speech reveals that there are two rather well-defined subproblems. The first is text-analysis, or alternatively linguistic analysis. The task here is to convert from an input text, which is usually represented as a string of characters, into a linguistic representation. This linguistic representation is usually a complex structure that includes information on grammatical categories of words, accentual or tonal properties of words, prosodic phrase information, and of course word pronunciation. The second problem is speech synthesis proper, namely the synthesis of a (digitally coded) speech waveform from the internal linguistic representation.

The artificial production of speech-like sounds has a long history, with documented mechanical attempts dating to the eighteenth century.[2] The synthesis of speech by electronic means is obviously much more recent, with some early work by Stewart [Stewart 1922] that allowed the production of static vowel sounds.[3] Two inventions, both from Bell Laboratories, helped initiate the modern work on speech synthesis that eventually led to the speech synthesizers that we know today. The first was the "Voder", by Homer Dudley, which was an electronic instrument with keys that allowed a skilled human operator to select either a periodic (voiced) source (for sonorant sounds), or a noise source (for fricative sounds); keys were also available for controlling a filter bank to alter the spectrum of the 'vocal tract' to approximate different vowel sounds, as well as some controls to produce stops; *fundamental frequency* (*F0*) was controllable by a foot pedal. Demonstrated at the 1939 World's Fair, the Voder clearly demonstrated the potential of the electronic production of speech. The second invention was the sound spectrogram [Potter *et al.* 1947], which for the first time gave a representation of speech that allowed researchers to discover robust acoustic correlates of articulatory patterns, and made it possible to conceive of automatically generating speech by rule. This second invention in particular was instrumental in the development of the so-called

[1] Of course, it is possible to hook a TTS system up to an optical character recognition system so that the system can read from printed text; but this does not affect the point, since the output of the OCR system will be a digitally coded text that serves as the input to the TTS system.

[2] For example, there was a mechanical device attributed to von Kempelen (1734–1804).

[3] Much of the historical data reviewed in this section is culled from Dennis Klatt's review article [Klatt 1987], which is highly recommended for anyone wishing to learn more about the history and problems of speech synthesis from its inception to the late eighties.

source-filter theory of speech production [Fant 1960], upon which most modern work on synthesis depends: simply put the source-filter theory states that speech can be modeled as one or more periodic or noise sources (the glottis or in the case of fricatives, a constriction), which are then filtered by an acoustic tube, namely the vocal tract (sometimes coupled with the nasal tract) from the point of the source onwards to the lips.

The three modern approaches to synthesis, namely *articulatory* synthesis, *formant* synthesis, and *concatenative* synthesis, have a roughly equally long history: the earliest rule-based articulatory synthesizers include [Nakata and Mitsuoka 1965, Henke 1967, Coker 1968]; the earliest rule-based formant synthesizer was [Kelly and Gerstman 1961]; and the concept of diphone concatenation was discussed in [Peterson *et al.* 1958] with the first diphone-based synthesizer reported being [Dixon and Maxey 1968]. We discuss these different approaches in somewhat more detail in the next section.

Text-analysis for speech synthesis has a much shorter history. One of the first systems for full text-to-speech conversion for any language — and certainly the best known — was the MITalk American English system. Although a full description was only published in 1987 [Allen *et al.* 1987], the work was actually done in the seventies. MITalk was capable not only of pronouncing (most) ordinary words correctly, but it also dealt sensibly with punctuation, numbers, abbreviations and so forth. MITalk was also significant in setting an important precedent: the approach taken to such problems as morphological analysis and grapheme-to-phoneme conversion was linguistically very well informed, with the approach modeled to a large extent on *The Sound Pattern of English* [Chomsky and Halle 1968]. The input of linguistic knowledge into TTS systems has remained high, and few people would seriously try to construct a full working TTS system based on a pure 'engineering' approach. Lamentably the comparable point cannot be said of speech recognition, where to this day linguistics has had little practical impact, and where the most sophisticated linguistic models incorporated in many systems are pronunciation dictionaries — which are often derived from a TTS system! Since MITalk, full TTS systems have been constructed for many languages, with several systems, including DECtalk, Lernhout & Hauspie, Berkeley Speech, Infovox and the Bell Labs system, being multilingual.

The purpose of this paper is to review the problems of text-to-speech conversion as it applies particularly to the conversion of Chinese text into Mandarin speech. Although we provide a synopsis of the literature in Section 3, most of this paper constitutes an overview of our own work on Mandarin synthesis at Bell Labs. There are a couple of reasons for this focus. First, while there is by now a sizable literature on Mandarin TTS, few systems are discussed in sufficient detail for us to be able to illustrate all of the issues involved in Mandarin TTS conversion by examples culled from the published literature. The second reason is that while no formal evaluations of our Mandarin system are available, informal assessments lead us to believe that it is among the best in terms of speech quality, and our (admittedly limited) knowledge of other systems leads us further to believe that our system's ability to deal reasonably with unrestricted text input is as good as that of any other system. Thus the Bell Labs Mandarin system is as reasonable a model as any upon which to base the discussion in this paper.

Before we begin examining issues pertaining particularly to Mandarin, however, it is necessary to review some basic concepts in speech synthesis, for readers who are not already familiar with the field of synthesis.

2 Review of Some Basic Concepts in Speech Synthesis

As noted above, there are three basic approaches to synthesizing speech, namely articulatory synthesis, formant synthesis and concatenative synthesis. Due to the complexity of the first, only the latter two approaches have so far proved commercially viable.

With one exception (waveform concatenation) all approaches to synthesis are based on the source-filter model introduced above. That is, the synthesis method can be broken down into two components, consisting of a model of the source and a model of the vocal tract *transfer function*. In normal speech, both the source and the transfer function change over time: the only exceptions to this are very unnatural utterances such as steady state vowels said on a constant pitch with no change in amplitude. It is standard to approximate the continuous variation in the source and filter, by dividing each second of speech into a set of a few hundred discrete steady-state frames. During synthesis, the source is filtered with the transfer function at each frame to produce the output speech parameters for that frame. Source models include models of the periodic vibration at the glottis during periods of voicing; and models of noisy supraglottal sources such as one finds with fricatives like *x* or *s*. Changes in the fundamental frequency of the voice (due, for example, to changes in phonological tones during the course of the utterance) are also handled by the glottal source model, by merely increasing or decreasing the rate of the pulses.

The main difference between the three basic methods of synthesis has to do with the way in which the sets of transfer functions for an utterance are computed. Articulatory synthesis attempts to solve the problem from first principles, as it were. Computational models of the articulators — the tongue, the lips, and so forth — are constructed that allow the system to simulate the various configurations that the human speech organs can attain during speech. Each configuration gives a particular shape to the vocal tract, and from these shapes, theories of acoustics are put to work to compute the transfer function for that particular shape. The difficulty is that both problems, namely mimicking human articulator motion, and computing acoustic properties from vocal tract shapes, are very hard. As a result, most work on articulatory synthesis has made simplifying assumptions about physical properties of the vocal tract, and has depended on two-dimensional mid-sagittal models. Only recently have researchers started seriously to build three-dimensional models of crucial components (such as the three-dimensional tongue model discussed in [Wilhelms-Tricarico and Perkell 1996]). Even with such models in place, however, our knowledge of articulatory phonetics is still sufficiently primitive that it will be many years before the articulators can be appropriately controlled using purely linguistic information, such as what one might compute from text.

Formant synthesis bypasses the complications of modeling articulator movements, and computing the acoustic effects of those movements, by computing spectral properties of the transfer function directly from the linguistic representation using a set of carefully designed rules. Consider a word such as 咖啡 *kāfēi* 'coffee'; Figure 1 gives a spectrogram for an utterance of this word by a male speaker of Beijing Mandarin; the formants (also called resonances) show up clearly as roughly horizontal dark bars in the spectrogram. For example, the vowel *a* has three formants in the lower half of the spectrogram between about 1.9 and 2.0 seconds. To synthesize this utterance using a formant synthesizer one needs to model as closely as possible the changes in the frequency and amplitude in the formants for the vowels *a* and *ei*, including the shape of the transitions of those formants into and out of the consonants, since these transitions are important in identifying the

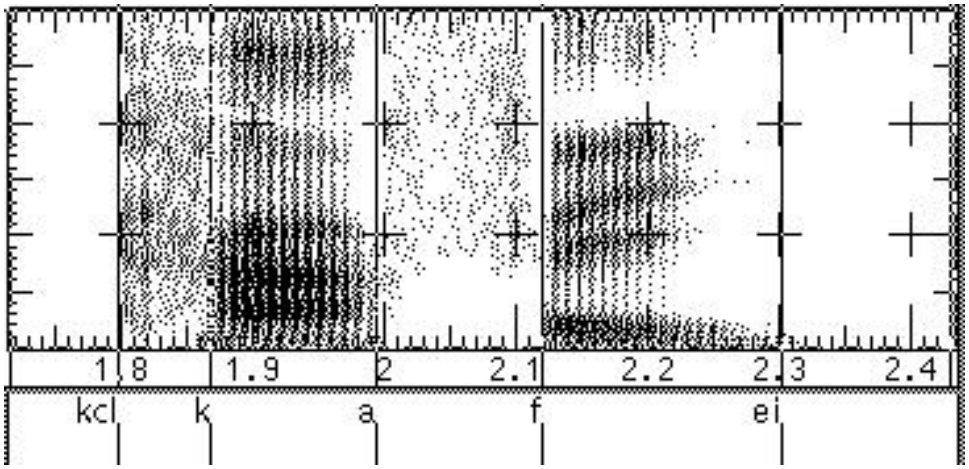

Figure 1: Spectrogram of an utterance of the word *kāfēi* 'coffee', by a male speaker of Beijing Mandarin. Note that 'kcl' denotes the closure portion of the *k*.

consonant. The transitions for *a* out of *k* are particularly clear, with a noticeable 'velar pinch' being present at the beginning of the *a* in the second and third formants, right before 1.9 seconds. Designing a set of rules to do this is certainly easier than constructing an articulatory model, but it is still a difficult task.

The simplest approach to synthesis bypasses most of these problems, since it involves taking real recorded speech, cutting it into segments, and *concatenating* these segments back together during synthesis. In the Bell Labs concatenative synthesis method, for example, recorded speech is first coded using *linear predictive coding* (LPC), which removes the source information, leaving one with a spectral representation of the time-varying transfer function of the utterance. Segments of this LPC representation are then cut out, based on cut points determined by examining a normal spectrogram, and stored in a table. The size and type of unit, whether it be a whole syllable, a half- or *demi*-syllable, phone-to-phone transition — a *diphone* — differs widely across systems, as we shall see more fully in the next section. Let us assume for the sake of discussion a diphonic system similar to the one used in the Bell Labs synthesis work. For an utterance such as 咖啡 'coffee' one would need the sequence of units [*–k] (where '*' represents silence), [k–a], [a–f], [f–ei] and [ei–*]. The [k–a] unit, for instance, would be selected starting during the silence just before the burst of the *k*, and extending to the right until just after the *a* reaches a steady state. In general, of course, a unit needed to synthesize a particular word will most likely *not* have been originally uttered as part of that word. So the [k–a] unit used for 咖啡 'coffee' might have come from 卡片 *kǎpiàn* 'card', and the [a–f] unit might have come from 麻煩 *máfán* 'troublesome'. The art in concatenative synthesis is to select units in such a way as to minimize discrepancies between adjacent units, which come originally from completely different utterances.

While many concatenative systems are like the Bell Labs system in that the source and filter are handled separately, the very notion of cutting up speech segments and reconcatenating them during synthesis naturally leads to the idea of bypassing the source-filter model entirely, and synthesizing directly in the *time domain*, that is by storing and concatenating waveform segments. In principle, the speech quality of such systems should be higher than that of LPC-based systems, in precisely

the same way as recorded speech is always of higher quality than LPC coded and resynthesized speech. This is not as easy as it sounds, since simply taking the two waveforms for the [k–a] of 卡片 'card' and the [a–f] of 麻煩 'troublesome' and concatenating them together will lead to very poor results due to the abrupt discontinuity that will be almost impossible to avoid. Furthermore, it is usually desirable to be able to change the fundamental frequency contour of the synthesized utterance from whatever contour was used for the utterance: one might want to change a low-toned utterance 打 *dǎ* 'hit' into a falling toned utterance 大 *dà* 'big'. This is also not trivial to do using a waveform representation. Nonetheless, the development of such techniques as *pitch-synchronous overlap and add* (PSOLA) [Charpentier and Moulines 1990] has allowed for the production of high quality speech synthesizers based on waveform concatenation. Still, waveform methods tend to be much more sensitive to precise placement of cut points than are LPC-based systems, and extreme variations in pitch remain a problem for waveform methods. One solution to the latter problem has been to simply record a given unit in all desirable tonal environments, thus giving a high degree of naturalness, but at the cost of a larger set of units.

This section has introduced concepts from just one part of the TTS problem, namely synthesis proper. Nothing has been said here about various approaches to text-analysis, of tone and intonation, or of segmental durations. We believe, however, that our discussion of these topics in ensuing sections will serve to illustrate the issues in those domains.

3 A Synopsis of the Literature on Mandarin Synthesis

In order to be able to place in a proper perspective the various projects on Mandarin TTS that have been and are being undertaken, it is useful to consider some of the more salient parameters that can be used to describe a TTS system. The following list is certainly not exhaustive, but it does hit the main points that are relevant for the Mandarin work:

1. What basic synthesis method is used? Is the system concatenative or rule-based?
   (a) If the system is concatenative, what basic coding scheme is used? Is the scheme time-domain (waveform, e.g. PSOLA) or frequency-domain (e.g. LPC)?
   (b) If the system is rule-based, do the rules describe formant trajectories, or is an articulatory model presumed?
2. Assuming the system is concatenative, what type of unit is used: are the units basically diphones, demisyllables, syllables, or some other unit?
3. How sophisticated are the prosodic models, in particular the models for duration and intonation? For example, is there any modeling of tonal-coarticulation — an important phenomenon in Mandarin phonetics?
4. Is the system a full *text*-to-speech system? By rights, a system can only call itself that if it is capable of converting from an electronically coded text written in the standard orthography for the language in question.

As we shall see, most Mandarin systems are concatenative, with an approximately equal split (as far as we can tell) between time- and frequency-domain coding schemes; most use syllables as

the basic unit; most early systems have fairly simple-minded prosodic modeling, with improved techniques becoming evident in later work; finally only since the very late eighties have full text-to-speech capabilities been reported.

3.1 Types of Synthesis and Size of Units

Concatenative synthesis. Mandarin synthesis work based on concatenative models dates to the early eighties, with the earliest examples being [Lei 1982, Huang *et al.* 1983, Zhou and Cole 1984, Lin and Luo 1985, Zhou 1986]. In this work, the emphasis was placed on the ability to fit the system into the tight speed and space constraints of the machines available at that time. As a result, the unit inventories tended to be small, with a favorite scheme being a pseudo-demisyllabic 'initial-final' (onset-rime) model: a syllable such as 開 *kāi* 'open' would be split into two units, namely the initial 'demisyllable' *k*, and the final 'demisyllable' *ai*. Such a scheme allows one to limit the number of units to around 60 [Huang *et al.* 1983, Qin and Hu 1988, Shi 1989], but only at a cost of naturalness, and even intelligibility, since no coarticulation effects *across* adjacent units are modeled. One other influence not modeled by these and many other kinds of concatenative inventories is the influence of tone. For example, there may be some correlation between vowel quality and the fundamental frequency of the voice, and this could lead to subtle problems if a particular unit that is recorded in one tonal environment is used to synthesize output for a different environment. Still, such tonal influences on a unit are generally believed to be quite small, certainly much smaller than the influence of abutting units. It is noteworthy, however, that one early system rather directly captures this tonal influence (while failing to capture the certainly more important cross-unit influence), by using a set of 184 pseudo-demisyllabic units collected from all possible tonal environments [Lin and Luo 1985].

Intra-syllabic coarticulation can be modeled better by changing to a true demisyllabic scheme. In the system described in [Mao *et al.* 1987], a set of 350 (non-tone-sensitive) demisyllables was used, consisting of 300 vowel-sensitive consonantal initial demisyllables, and 50 vocalic final demisyllables. In such a scheme, a syllable such as 開 *kāi* 'open', would be synthesized by the two units *ka* and *ai*. Another system reported in [Chiou *et al.* 1991] uses 93 vowel-sensitive initial demisyllables. [4] Using demisyllables is clearly preferable over the earlier pseudo-demisyllabic scheme in that one can now model the influence of the vowel on the consonant initial, in a way that was not possible in the earlier scheme.

A logical endpoint of this progression of ever larger and more context-sensitive units are the various syllable-based concatenative systems [Ouh-Young 1985, Ouh-Young *et al.* 1986, Liu *et al.* 1989, Liu *et al.* 1991, Ju *et al.* 1991, Chan and Chan 1992, Cai *et al.* 1995, Chu and Lü 1995, Hwang *et al.* 1996]. Syllable-sized units needless to say allow one to perfectly model within-syllable coarticulation. However, they merely shift up one level the problem faced by pseudo-demisyllabic systems in that merely having an inventory covering all of the syllable types of Mandarin does not allow one to model coarticulation effects that cross syllable boundaries. As a result, many of these systems must resort to inserting short silences between adjacent syllables to avoid other audible artifacts that would arise if the syllables were directly concatenated. But this in turn results in an unnatural 'choppiness', or even lowered intelligibility in some instances. One instance where one might expect inserted silences to cause problems is in medial syllables that begin with fricatives.

[4] One unique characteristic of this system is that the vowels are generated by a formant synthesizer.

Part of the acoustic distinction between a fricative such as *sh* and an affricate such as *ch* is that the latter begins with a silence. So it is likely that a fricative in a sequence like *n $ sh* (where '$' marks a syllable boundary) will be confused with an affricate (*n $ ch*) due to the inserted silence, especially if the silence is sufficiently long: this would lead to the misinterpretation of 人少 *rén shǎo* 'there are few people' as 人吵 *rén chǎo* 'people are noisy'. Indeed, this is precisely what was observed in [Zhang *et al.* 1995], wherein it was noted that the most frequent confusion in several waveform-based Mandarin systems was between voiceless fricatives, and aspirated plosives or affricates. The cause, we suspect, was not the waveform-based coding schemes themselves, but rather the silences that were inserted between syllables in what almost certainly were largely syllable-based systems.

There is at least one attempt to build a word-based system [Xu and Yuan 1993], where syllables as well as frequent words are collected in the speech database, which contains around 10,000 units. The problem facing such systems is of the same nature as the syllable-based system: there is no way to model coarticulation effects between words.

With respect to type of analysis method used, linear-predictive and waveform-based approaches are roughly evenly balanced in the literature, with a slight preference for the former. LPC systems include [Huang *et al.* 1983, Lin and Luo 1985, Ouh-Young *et al.* 1986, Mao *et al.* 1987, Qin and Hu 1988, Liu *et al.* 1989] as well as the Bell Labs systems. PSOLA systems include [Chu and Lü 1995, Hwang *et al.* 1996, Choi *et al.* 1994, Cai *et al.* 1995].

Rule-based Synthesis. Rule-based (formant) synthesizers for Mandarin have an earlier vintage than concatenative systems, with the earliest citation that we have been able to find being [Suen 1976]. Suen used a VOTRAX synthesizer to produce Mandarin speech from a phonetic transcription, coupled with five tonal contours in simple citation-form. Intelligibility was evidently a problem (as it was for any synthesizer at that time): "subjects could hear the synthesized passage with an encouraging intelligibility score of over 70% after less than an hour of exposure to the sounds." Subsequent work in a similar vein is [Lee *et al.* 1983]. Later rule-based synthesis work includes work at KTH [Zhang 1986], Matsushita [Javkin *et al.* 1989], Institute of Acoustics, Academia Sinica (Beijing) [Lü *et al.* 1994], Taiwan University [Shie 1987, Hsieh *et al.* 1989], and the Chinese Academy of Social Sciences (Beijing) [Yang and Xu 1988]. The latter two systems share the property of being syllable-based in a rather peculiar sense: rules are used to generate parameters to synthesize single syllables, and these syllables are then *concatenated* together, much as in a purely concatenative system. In neither system are there rules to handle cross-syllable coarticulation.[5]

Synopsis. So, all of the concatenative work, and at least some of the rule-based work, that we have discussed so far can be described as syllable-based in that the basic units of the system either *are* syllables simpliciter, or else are some well-defined constituents of a syllable, where if any cross-unit coarticulation is modeled, it is only *syllable-internal* coarticulation. As we have argued, this predisposition towards syllable-based systems is problematic from the point of view of producing natural-sounding speech. One solution to this problem is to abandon syllable-based models in favor of units which can model both syllable-internal and cross-syllable coarticulation. One such unit is

[5]To be fair, we should note that [Yang and Xu 1988] represents work that was presented by the authors themselves as incomplete: we believe that in subsequent work, cross-syllable dependencies have been modeled, though we have been unable to find a citation to such later work.

the diphone. In a purely diphonic system, a disyllabic sequence such as 開門 *kāi mén* 'open the door' might be constructed out of the units [k–a ai–m m–e e–n]. Greater or lesser quality can be obtained by contextualizing the diphonic units. For example in the 1987 Bell Labs female-voice Mandarin synthesizer [Shih and Liberman 1987], the *k-a* unit used in synthesizing 開 *kāi* 'open' was the same as the unit used in synthesizing the *kā* of 咖啡 *kāfēi* 'coffee.' But this is not completely satisfactory since the *a* in *kāi* is higher and more fronted than the *a* in *kā*. Thus, in the new Bell Labs Mandarin synthesizer, presented in more detail below (and see also [Ao *et al.* 1994]), the [k–a] units used for these two syllables are different. Apart from our own systems, the only other system that we are aware of which is diphone-based is the Apple system, described in [Choi *et al.* 1994].

3.2 Models of Prosody

All synthesis systems, strictly speaking, have a prosodic model in the sense that fundamental frequency, duration, and intensity must be assigned in the production of speech. An extremely simple system for Mandarin might assign to each syllable a tone shape selected by the lexical tone, and assign constant duration and constant intensity to each phone. The prosodic model in early Mandarin synthesis systems did not go much beyond this simple description. Later systems have fairly sophisticated models predicting variations of tone and duration suitable for the context. However, it is fair to say that there are at this point in time no text-to-speech systems for any language that incorporate fully adequate models of prosody.

There are many causes which contribute to the difficulty of prosodic modeling. Even for the plain, non-expressive reading style, there are many factors controlling prosody. There are also immense variations among speakers, and even between different renditions of the same text by the same speaker: these considerations make it hard to discover the robust underlying regularities governing prosody. Modeling of expressive reading styles is even harder, especially since no TTS system can be said to truly understand what it is reading: thus important information, such as when to emphasize a word and how much emphasis to put in, cannot generally be reliably predicted.

Models of Duration. Few papers on Mandarin synthesis systems report on duration modeling. In early systems, most of the effort was channeled into building the acoustic inventory and, to some extent, into constructing models of tone. Among the relatively few published reports on duration modeling, several different approaches are taken. The systems reported in [Shih and Liberman 1987, Chiou *et al.* 1991, Choi *et al.* 1994] use hand-crafted duration rules based on published reports such as [Ren 1985, Feng 1985], along with some educated guesses to bridge the gaps not covered in the published phonetic work on duration. In contrast, the new Bell Labs Mandarin system [Shih and Ao 1994, Shih and Ao 1996] (described more fully in Section 7), and the Chiao-Tung University system [Hwang and Chen 1992, Hwang *et al.* 1996] derive duration values from a labeled speech database. The Bell Labs model is a parametric mathematical model, the parameters of which are trained based on duration values in the database; the Chiao-Tung system is based on neural networks. In our own experience, there is drastic improvement in our current duration model over the 1987 rule-based system, and as a result, both the intelligibility and naturalness of synthesized speech improves greatly. The contrast of the two systems attests to the importance of a good duration model.

The results of the Chiao-Tung neural-network-based system are good, especially in contrast

to a simple rule-based system. And the advantages of using neural nets are familiar: in general, one can achieve passable behavior by training the network on sufficient amounts of appropriately labeled training data, without having to do any analysis beforehand of what kinds of factors are likely to be relevant, and what their likely interactions are. This is obviously much less work than a sophisticated hand-crafted rule-set, such as the well-known duration model for the MITalk system reported in [Allen *et al.* 1987], which depend upon a great deal of phonetic understanding on the part of the rule writer. And a neural net approach is also less onerous than a mathematical model of the kind we describe in Section 7. However, there are various reasons for preferring the approach we have chosen, over a neural net approach. The main reason is that general statistical modeling methods, such as neural nets or CART [Breiman *et al.* 1984], suffer from sparse data problems, in that they are often unable to produce reasonable results in cases involving combinations of factors that were unseen in the training data; in contrast, the duration model we present below, is much more robust in the face of such unseen factor combinations.

Models of Tone and Intonation. The simplest tone model for Mandarin synthesizers just specifies the F0 contours of Mandarin lexical tones [Suen 1976, Lin and Luo 1985, Qin and Hu 1988, Mao *et al.* 1987]. It was soon realized that even a few simple tonal coarticulation rules improve the smoothness of the synthesized speech. Many systems implemented the rules described in [Chao 1968], including the neutral tone rule, tone sandhi rules, the half tone 3 rule, and the half tone 4 rule [Zhang 1986, Ouh-Young *et al.* 1986]. Some systems also include tone rules that are intended to capture the tendency for pitch to drop as the sentence proceeds: for example, a tone 1 following a tone 3 is replaced by a variant of tone 1 with lower pitch [Lee *et al.* 1989, Liu *et al.* 1989, Chiou *et al.* 1991, Lee *et al.* 1993].

A system built at University College London [Shi 1989] incorporated the intonation model of [Gårding 1987], where intonation effects are represented by *grids* which define the global F0 range and shape. Tones are then realized within the confines of the grids. This system does not have the capability of predicting intonation effects automatically and requires annotated input. The system reported in [Zhou and Cai 1995] also allows hand-annotation for stress.

The system discussed in [Chu and Lü 1995] follows the model of [Shen 1985], where the intonation *topline* reflects the level of emphasis, with higher values corresponding to stronger emphasis. The *baseline* is considered to reflect the strength of syntactic boundary, with lower F0 value corresponding to stronger boundaries.

A few systems attempt to capture tone and intonation variation using statistical models [Chang *et al.* 1991a, Chang *et al.* 1991b, Chen *et al.* 1992a, Wang *et al.* 1992] including neural networks [Chen *et al.* 1992b, Hwang and Chen 1994, Hwang and Chen 1995, Wu *et al.* 1995, Chen *et al.* forthcoming]. In order to facilitate clustering analysis on tones, these systems typically represent tones as a four-dimensional feature vector obtained by four orthonormal polynomials [Chen and Wang 1990], where the first coefficient represents the F0 mean and the other three represent the tone shape. After clustering analysis, various numbers of representative tone patterns are chosen and stored in codebooks. The models are trained to match tone patterns with selected contextual features. As with the training of duration models, these statistically-based methods are capable of creating an impressive prosodic model in a short time without linguistic analysis. However, the problem of unseen factor combinations, which we saw in the discussion of duration modeling, is even more serious in the modeling of F0. This is evident in the report of [Chen *et al.* 1992a]: while

77.56% of the tone patterns chosen by the model match that of the natural speech in the training data, (a very impressive score compared with the ceiling of 82.96%, obtained by comparing different repetitions of the speech database), the performance declines drastically to 40% for the test set.

The Bell Labs system is similar to the systems of [Zhang 1986, Ouh-Young *et al.* 1986] in using a rule-based tone and intonation model, but there are some notable differences. Following [Pierrehumbert 1980, Liberman and Pierrehumbert 1984], tones are represented as abstract targets, and sentential effects are separated from tonal coarticulation rules. We found experimentally that the sentential lowering effect primarily affects tone height, not tone shape, and the amount of lowering can be modeled as an exponential decay. This process can in principle create an infinite number of tone-height distinctions, which is not something that can be handled by tonal replacement rules. We also carried out phonetic studies involving all possible tonal combinations and derived a set of tonal coarticulation rules that is more complete than previously reported results [Shih 1987]. A crucial aspect of our tonal analysis is the strict separation of *phonological* tone sandhi rules from *phonetic* tonal coarticulation rules. These two tonal phenomena seem to be frequently confounded in the literature, but it is important to keep them distinct. For example, in the statistical approach discussed in [Chen *et al.* 1992a] no distinction is made in the training data between surface tone 3 (e.g. 買 *mǎi* 'buy') versus a *sandhied* tone 3 (e.g. 買酒 *mai[2] jiǔ* 'buy wine'),[6] which is identical to tone 2. This results in redundancy in that all the tone shapes for tone 2 must also be stored in the codebook under the entry for tone 3. Furthermore, while the statistical model does capture local tone sandhi changes, more or less by brute force, it is possible to construct reasonable sentences where the realization of an underlying tone 3, which precedes another underlying tone 3, depends upon information that is many syllables to the right in the sentence. Such cases could not be readily handled using a purely statistical approach that confounds phonological and phonetic tone changes. We discuss these examples, and other details in the phonological analysis of Mandarin third tone sandhi in Section 5.3; we further discuss phonetic tonal rules in Section 8.

3.3 Text-analysis Capabilities

Most early systems for Mandarin presumed an input in some form of phonetic transcription, such as pinyin. For example, the 1987 Bell Labs system allowed input in pinyin, optionally annotated with bracketings indicating syntactic constituent structure. Such systems do not qualify as text-to-speech systems in the strict sense, since they do not allow input represented in the normal orthography of the language.

The first system that contained at least some text-analysis capabilities and was thus able to convert from normal Chinese orthography into Mandarin speech was the Telecommunications Labs system [Liu *et al.* 1989]. This system was eventually linked to an optical character recognition (OCR) system for converting printed text into sound.[7]

Of course, by the mid nineties there were several Mandarin systems that could read from (electronically coded) Chinese text, including the Apple system [Choi *et al.* 1994], several Mainland systems (e.g. [Cai *et al.* 1995]) and our own system, which we describe in more detail below. There remain, however, relatively few publications on the text-analysis problems for Chinese TTS,

[6] We use the notation '[2]' to denote a sandhied tone 3.

[7] Such systems have been available for English for many years; see for example [Kurzweil 1976]. But the OCR problem in English is not nearly as formidable as it is in Chinese.

with our own publications in this area being arguably the most extensive. This dearth of publications mirrors the lamentable second-class status which text-analysis matters have in the worldwide synthesis community. Although the word *text* is one of the two content words in the phrase *text-to-speech*, the amount of published material in this area comprises significantly less than half of the published material on synthesis as a whole.

3.4 Summary

The nineties has seen a large increase in the number of reported Mandarin synthesis systems, with several systems being improved versions of previous systems, as well as many completely new systems. Speech quality has improved dramatically, both in terms of segmental quality, and in terms of more natural sounding prosody. Several newer systems are full text-to-speech systems in the sense we have defined. Finally, there has been a small amount of work on synthesis for other Chinese languages: for example Taiwanese synthesizers have been built at Tsing Hua University [Wang and Hwang 1993], Bell Laboratories and Telecommunication Labs; a Cantonese synthesizer has been built at Hong Kong University [Lo and Ching 1996].

A summary of the Mandarin text-to-speech systems is given in Table 1, listing the most salient attributes of these systems. Demo sentences of some of these systems will be included in the Bell Labs TTS web page currently under construction. The address will be http://www.bell-labs.com/project/tts/.

There has been very little work on formal evaluation of Mandarin synthesizers, with the exception of the Mainland assessment program discussed in [Lü *et al.* 1994, Zhang *et al.* 1995, Zhang 1996], where systems are judged on the base of syllable, word, sentence intelligibility, as well as overall naturalness. Systems with text analysis capability are further evaluated on their performance on selected texts. Eight systems were evaluated in the 1996 report.

3.5 The syllable as a basic unit of synthesis: a critique.

As we have seen, the vast majority of Mandarin synthesizers are based on syllable-sized units. In this regard they differ sharply in design from, for example, our own Mandarin synthesizer. Rather than merely noting this distinction however, it is instructive to ask why it is such a prevalent notion that Mandarin synthesizers should be based on syllable-sized units, rather than units of some other size.[8]

There seem to be two reasons for this prevalence. The first is that as a practical matter, the number of syllables in standard Mandarin seems sufficiently small – around four hundred, without counting tonal distinctions – that one can easily conceive of building an inventory of all syllable types; even if one allows for tones, one would only need to deal with about fourteen hundred types. For English, of course, a syllable-based approach would be far less palatable: a dictionary of somewhat over a hundred thousand words yields approximately fifteen thousand syllable types, and this almost certainly does not exhaust the set of possible types. But while the contrast seems

[8] Note that it is not only in the published literature where one finds a strong bias towards syllable-based synthesis. On several occasions in informal discussions with Chinese speech researchers we have been told that syllable-based systems are clearly the correct approach. The prejudice is a widespread one indeed. The primacy of syllables is also evident in work on speech recognition for Mandarin.

Inst. or author	Synth. Meth.	Unit	Text Anal.	Prosody	date
Suen	VOTRAX	phoneme	no	5 tones	1976
Lee et.al.	VOTRAX	phoneme	no	5 tones	1983
Huang et.al.	LPC	pseudo-demisyllable	no	4 tones	1983
Zhou et.al.		pseudo-demisyllable	no	7 tones	1984
Lin & Luo	LPC	toned phoneme	no	no	1985
Taiwan U.	LPC	syllable	no	rules	1985
KTH	formant		no	rules	1986
Bell Labs	LPC	diphone	no	rules	1987
Taiwan U.	formant	syllable	no	rules	1987
Tsinghua (Beijing)	LPC	demisyllable	no	4 tones	1987
Acad. of Soc. Sci.	formant	syllable	no	stat. model	1988
Qin & Hu	LPC	pseudo-demisyllable	no	4 tones	1988
U.C. London	formant	pseudo-demisyllable	no	rules	1989
Matsushita	formant	pseudo-demisyllable	no	model	1989
Telecom. Labs	LPC	syllable	yes	rules/stat. model	1989
Tsing Hua U.	hybrid	demisyllable	no	rules/stat. model	1991
Telecom. Labs	LPC	toned syllable	yes	rules	1991
Chiao Tung U.	PSOLA	syllable		stat. model	1992
Chiao Tung U.	LPC	syllable		stat. model	1992
Hong Kong U.	LPC	syllables	no	rules	1992
Xu et.al.		words	yes	rules	1993
Bell Labs	LPC	diphone	yes	rules/stat. model	1994
Inst. Acoustics	formant		yes	rules	1994
Apple	Apple	diphone	yes	rules	1994
Tsinghua (Beijing)	PSOLA	toned syllable	yes	rules	1995
Inst. Acoustics	PSOLA	toned syllable	yes	rules	1995
Cheng Kung U.	CELP	toned syllable	yes	stat. model	1995

Table 1: Mandarin TTS Systems

plain enough, it is misleading for a couple of reasons. Firstly, the count of four hundred types is certainly an undercount, when one considers new syllable types created as a result of 兒 *er* suffixation in northern dialects, and the various reduced syllables that one encounters in fluent Mandarin speech. Thus one finds that 他們 *tāmen* 'they', which would canonically be transcribed as *tāmən*, is often pronounced *tām*, which is not a legal syllable according to the normal inventories of Mandarin. Certainly one may with some legitimacy exclude such reductions in a *text*-to-speech system on the basis of the fact that they do not represent reading style. But assuming that the ultimate goal of speech synthesis is to produce natural-sounding speech of all styles, synthesis of Mandarin on the basis of a fixed inventory of four hundred syllables is bound to fail. Secondly, syllable-based *concatenative* systems still have to deal with the problem of cross-syllable and cross-word coarticulation, as we have already seen. In a purely concatenative system it would certainly be theoretically possible to store contextually determined syllable-sized units, but the number of such units would inevitably be huge, thus defeating the numerical argument for using syllables that was advanced in the first place.

The second reason for the prevalence of syllable-based systems relates to a much deeper issue, we believe. For many centuries Chinese philological tradition has centered on the written character (字); larger linguistic units had little formal status. The reason for this is clear in that characters most commonly correspond to single morphemes, and vice versa (though see [DeFrancis 1984] for some refinement of this point). Furthermore, they are graphically rather distinct, thus making the morphological analysis of a segment of text transparent in a way that has no parallel in, say, European languages. In analyzing Chinese, there was thus a very strong incentive to view the character/morpheme as the minimal unit of the language, much as Bloomfield [Bloomfield 1933] later analyzed the morpheme as the minimal unit of meaning in all languages. Since characters, with few exceptions, correspond to single syllables, it is a short step to viewing the syllable in Chinese as being the basic unit of phonological analysis. However, a wealth of phonological evidence for subsyllabic units in Mandarin, in addition to the more practical problems with using syllables that we have already discussed, suggest that this step is misguided.

PART TWO: THE BELL LABS MANDARIN TTS SYSTEM

4 Overall Structure of the TTS System

The Bell Labs TTS system consists of about ten modules which are pipelined together. With the exception of the American English system, which was built much earlier and has a somewhat different structure, the software is multilingual in that none of the programs contain any language-specific data: thus the set of programs that we use for Mandarin have also been used for other languages, including German, Spanish, Russian, Romanian and Japanese.[9] All that is necessary to switch languages is to provide an appropriate set of language-specific tables for text-analysis, duration, intonation and concatenative units.

[9]This statement is slightly inaccurate in that we currently have two competing models of intonation. One is based on the Liberman-Pierrehumbert-style tonal target model, which we use for Mandarin and describe below. The other one, which is more similar to the Fujisaki model, is used for our set of European languages, none of which are tonal or pitch-accent languages. These involve two different intonation modules.

The ten modules are as follows:

1. The **text-analysis** module *gentex*.
2. The **phrasing** module, which simply builds phrases as specified by gentex.
3. The **phrasal phonology** module, which sets up phrasal accents.
4. The **duration** module.
5. The **intonation** module.
6. The **amplitude** module.
7. The **glottal source** module.
8. The **unit selection** module, which selects the concatenative units from a table, given the phonetic transcription of the words.
9. The **diphone concatenation** module, which performs the actual concatenation of the units.
10. The **synthesis** module, which synthesizes the final waveform given the glottal source parameters, the F0 values, and the LPC coefficients for the diphones.

In this paper, we will discuss the Mandarin-specific data that relates to modules 1, 4, 5, 8 and 9.

With the exception of the unit-concatenation module, which outputs a stream of LPC parameters and source state information for the waveform synthesizer, and the synthesis module itself, which reads the LPC and source-state stream, each module in the pipeline communicates with the next via a uniform set of data structures. Information is passed along on a sentence-by-sentence basis, and consists of a set of arrays of linguistic structures. Each array is preceded by the following information:

1. The structure's type T: e.g., word, syllable, phrase, phoneme, concatenative unit ...
2. N, the number of structures of type T for this sentence
3. S: the size of an individual structure of type T

After each of these headers is an $N \cdot S$ byte array consisting of N structures of type T.

There are a number of advantages to this modular pipeline design. The first advantage is a standard observation: if the system's modules share a common interface then it is relatively easy for several people to work on different modules independently of one another, as long as they agree on the input-output behavior of each module, and the manner in which the modules should relate to one another. Second, the pipeline design makes it easy to cut off the processing at any point in the pipeline. During the design of a segmental duration module, for example, one might want to test the system by running it over a large corpus of text, and sampling the segmental durations that are computed. For that task one would not normally care about what happens in modules subsequent to duration, such as intonation, amplitude or dyad selection. In the Bell Labs TTS architecture, one can simply run all modules up to and including the duration module, and then terminate the pipeline with a program that prints out information about segments and their durations. Similarly, it is easy to insert into the pipeline programs that modify TTS parameters in various ways.

5 Text Analysis

We will describe the overall framework for text-analysis only briefly here, concentrating rather on giving some more specific details that are relevant to Mandarin. A more detailed discussion of the model is given in [Sproat 1995, Sproat 1996]; an in-depth analysis of the approach as it applies to Chinese word-segmentation is given in [Sproat *et al.* 1996].[10] While the discussion here naturally focuses on Chinese, the reader should bear in mind that the same model has been applied in synthesizers for seven other languages besides Mandarin: Spanish, Italian, French, Romanian, German, Russian and Japanese.

5.1 Lexical Models

To see how the text-analysis module works, let us start with an example sentence, namely 史立璿有2隻老鼠。 'Shi Lixuan has two mice.'. We presume that the correct segmentation of this sentence into words is 史立璿 (Shi Lixuan) 有 (has) 2 (2) 隻 (classifier) 老鼠 (mouse) and 。, and that the pronunciation should be *shǐlìxuán you[2] liǎng zhī lao[2]shǔ*. One can view a sentence (or a text generally) as consisting of one or more words coming from a model, separated by word delimiters. In English, word delimiters include whitespace, as well as punctuation. In Chinese, of course, space is never used to delimit words. We can model this formally by representing the interword 'space' in Chinese with the empty string ϵ; this is then used, in addition to possible punctuation marks, as a delimiter of words. Each word model maps words in the text that (possibly) belong to that model, to a lexical analysis of that word. So a personal name model maps the string 史立璿 to the lexical analysis 史立璿{urnp}, which simply tags the string as an unregistered proper name; an ordinary word model maps 有 to 有{vb}; a model of numbers maps '2' into 兩{num} *liǎng*; finally the ordinary word model maps 隻 into 隻{cl}, and 老鼠 into 老鼠{nc} (where *nc* means 'common noun'). Of course, precisely because Chinese does not generally delimit words in text, most sentences contain potential ambiguities in word segmentation. In the particular sentence at hand, there are no other plausible segmentations than the one being considered, but it is nonetheless the case that 老 *lǎo* 'old' and 鼠 *shǔ* 'rat' are both words in our dictionary, so it is at least possible to segment the text in a way that splits these two characters into separate words. One way to deal with this ambiguity is to assign costs or weights to the different possible analyses. As described in more detail in [Sproat *et al.* 1996], individual words in the lexical models are assigned a cost that is equal to the estimate of the negative log unigram probability of that word occurring. For common words such as 老鼠 it is possible to estimate this cost directly from a corpus. For unseen words, including unseen personal names like 史立璿, one must estimate the cost using a model: for personal names we follow a model proposed in [Chang *et al.* 1992] with some modifications as discussed in [Sproat *et al.* 1996].

The lexical models are implemented as *weighted finite-state transducers* (WFST) [Pereira *et al.* 1994, Pereira and Riley 1996], which are constructed using a lexical toolkit described in [Sproat 1996]. For ordinary words, including morphologically derived words, a weighted finite-state morphological model is constructed which simply lists all words in the lexicon along with words that can be derived with productive affixes (such as the plural marker 們 *men*). Weighted finite-state

[10] Though the latter paper describes the problem in a slightly different way from the way it is described here. Note that the cited paper also gives a fairly extensive performance analysis for the segmentation.

models are also constructed for personal names, and foreign names in transliteration. For example, personal names are allowed to follow the pattern FAMILY+GIVEN, where FAMILY names belong to a small list of a few hundred names, weighted by their frequency of occurrence, and GIVEN names can be basically any character, weighted with an estimate of its likelihood in a name.

The construction of the numeral expansion transducer requires a little more explanation. The construction factors the problem of expanding numerals into two subproblems. The first of these involves expanding the numeral sequence into a representation in terms of sums of products of powers of the base — usually ten; call this the *Factorization* transducer. The second maps from this representation into number names using a lexicon that describes the mapping between basic number words and their semantic value in terms of sums of products of powers of the base; call this the *Number Lexicon* transducer. A sample Number Lexicon fragment for Mandarin is shown in Figure 2. Note that unlike the case of ordinary words and names, the weights here — subscripted '1.0' for some entries — are arbitrary, since one does not in general need to have an estimate of the likelihood of a sequence like *23*, in order to be sure that it should be treated as a unit; but whereas the mapping from {2} to 二 *èr* is free, the mapping from {2} to 兩 *liǎng* has a positive cost, which means that in general this will be a disfavored expansion. The entry 三$_1$ indexes the particular pronunciation of 三 — in this case *sān* — that should be used (another possible pronunciation is *sàn*. Similarly, 一$_1$, 一$_2$, 一$_3$ index, respectively, the pronunciations *yī*, *yí* and *yì*, each of which is appropriate in a different context: see below on how these different pronunciations are selected depending on the context.

The first stage of the expansion is language- or at least language-area dependent since languages differ on which powers of the base have separate words (see [Hurford 1975], inter alia): so Chinese and Japanese have a separate word for 10^4, whereas most European languages lack such a word. The full numeral expansion transducer is constructed by composing the Factorization transducer with the transitive closure of the Number Lexicon transducer. The construction of the numeral expander for Mandarin is shown schematically in Figure 3.

Additional Mandarin-specific cleanup is required for numeral expansion over and above what was said here. For example, the lexicon given in Figure 2 coupled with the model just described would predict *二百 *èrbǎi* instead of 兩百 *liang[2]bǎi* for *200*, and *一$_1$千００一$_1$ *yīqiān líng líng yī* rather than 一$_3$千０一$_1$ **yì***qiān* **líng** *yī* for *1001*. This cleanup can be accomplished by rewrite rules such as those given below, which guarantee that the 一$_3$ allomorph of 一 is used before 百 'hundred' and 千 'thousand'; that 兩 is similarly used instead of 二 in the same environments; and that strings of ０ reduce to just one ０:[11]

０$^+$	→	０	/	Number ___ Number
二	→	兩	/	___ 百|千
一$_1$	→	一$_3$	/	___ 百|千

Other rules select the allomorphs 一$_2$ for 一 and 兩 for 二 before 萬 *wàn* 'ten thousand' and 億 *yì* 'hundred million'. As shown in [Kaplan and Kay 1994], given certain conditions, sets of rewrite rules of this kind can be compiled into finite-state transducers. The particular algorithm for achieving this compilation that is used here is described in [Mohri and Sproat 1996].

[11] The 二/兩 allomorphy rule is of course just a special instance of the measure word-induced allomorphy discussed below.

$$
\begin{array}{lcl}
/0 & : & \text{〇}/ \\
/1 & : & \text{一}_1/ \\
/1 & : & \text{一}_{2\ 1.0}/ \\
/1 & : & \text{一}_{3\ 1.0}/ \\
/2 & : & \text{二}/ \\
/2 & : & \text{兩}_{1.0}/ \\
/3 & : & \text{三}_1/ \\
/4 & : & \text{四}/ \\
 & \vdots & \\
/10^1 & : & \text{十}/ \\
/10^2 & : & \text{百}/ \\
/10^3 & : & \text{千}/ \\
 & \vdots &
\end{array}
$$

Figure 2: Mandarin number lexicon. The symbol sequence on the left-hand side of each row maps to the symbol sequence on the right-hand side. See the text for an explanation of the symbols 一_1, 一_2, 一_3 and 三_1. Subscripted costs of 1.0 mark some entries as being disfavored.

$$
\begin{array}{lcll}
334 & \circ & \boxed{\text{Factorization}} & \Rightarrow \quad 3 \cdot 10^2 + 3 \cdot 10^1 + 4 \\
 & \circ & \boxed{\text{NumberLexicon}}^* & \Downarrow \\
 & & \text{三}_1\text{百三}_1\text{十四} &
\end{array}
$$

Figure 3: Expansion of *334* in Mandarin.

Returning now to our example, recall that we defined a sentence as consisting of one or more words separated by word delimiters. We can now define this somewhat more formally as follows. Call the ordinary word WFST W, the personal name WFST P, and the numeral expansion WFST N. Then a *single* text word can be mapped to all possible lexical analyses by the union of these transducers (along with other lexical transducers), or in other words $W \cup P \cup N$. The interword delimiter model for Chinese simply maps ϵ to a word boundary # — denoted as $\epsilon : \#$, and punctuation marks such as 。 into various types of prosodic phrase boundary: let us denote this as $punc : \phi$, for ϕ a phrase boundary. For example we assume that period 。 maps to an *intonational* phrase boundary I.[12] (As we note below, this presumes that punctuation *always* signals a prosodic phrase boundary, whereas a prosodic phrase boundary *never* occurs between two adjacent words: clearly this is wrong, and this is something that we plan to improve upon in future work.) The interword model I is thus given as $(\epsilon : \#) \cup (punc : \phi)$. Finally, a sentence can be defined as $((W \cup P \cup N)^\frown I)^+$, or in other words one or more instances of a word separated by interword

[12] We assume some familiarity on the part of the reader with the concept of *prosodic phrase*. A particularly clearly articulated theory of prosodic phrasing is described in [Pierrehumbert and Beckman 1988]

material: call this the *lexical analysis WFST* L. Representing the input sentence as an unweighted acceptor S, we can then *compose* the sentence with L to yield the weighted lattice $S \circ L$ of all possible lexical analyses. One could then apply a Viterbi search to this lattice to find the best analysis based purely upon the lexical costs.

5.2 Language Models

However, it is often desirable, as in the case of the example sentence at hand, to do some analysis of context in order to more finely hone the analyses represented in the lattice, and possibly eliminate some from further consideration. Adopting the standard terminology from speech recognition, we term such contextual models as *language models*. The contextual model in the current Mandarin system is rather simple, dealing only with the allomorphy of numerals before classifiers. When 'two' occurs as a simple numeral (i.e., *not* in a compound numeral like 'twenty two') before a classifier such as 個 *gè*, 隻 *zhī* or 條 *tiáo*, it should take on the form 兩 *liǎng* rather than 二 *èr*. This can be accomplished by a filter that simply disallows 二 occurring as an expanded numeral (tag {num}), directly before a classifier. An expression of the following form where HANZI stands for any character, and Σ is (conventionally) any label in the alphabet of the transducer, accomplishes this:
$\neg(\Sigma^*\#二\{\text{num}\}\#\text{hanzi}^+\{\text{cl}\}\Sigma^*)$

Represented as a transducer T, and composed with $S \circ L$, the language model will weed out analyses that violate the constraint: in the example sentence, analyses including the sequence 二 隻 are ruled out. Now $BestPath(S \circ L \circ T)$ yields the desired result: #史立璿{urnp}#有{vb}#兩{num}#隻{cl}#老鼠{nc}I. Note that this analysis includes the information that the sentence is a single prosodic phrase, since the only phrase boundary is the marker I, mapped from 。.

5.3 Phonological Models

Having singled out a lexical analysis, it is then necessary to produce a phonological analysis. For the most part, this can be handled by a simple transducer that maps from the lexical representation of the characters into their phonemic representation: note that homographs such as 咖 — *kā* in 咖啡 *kāfēi* 'coffee', but *gā* in 咖哩 *gālǐ* 'curry' — are generally disambiguated as part of the lexical analysis (though see below). The two main complications in the transduction to phonetic representation involve two rules of tone sandhi, namely third tone sandhi and the special sandhi rule for 不 'not' whereby the default fourth-tone variant *bù* is changed to a rising tone variant *bú* before a following falling tone.[13]

Third Tone Sandhi. Consider first third tone sandhi. This rule, which is described in detail in [Shih 1986] changes a (low) tone 3 into a (rising) tone 2 when it precedes another tone 3. We notate this derived tone 2 as '[2]':
$3 \rightarrow [2] \; / \; \_\_\_ \; 3$
The rule applies cyclically, applying to the smallest prosodic structure first, and to successively

[13] In a rule remiscent of the rule for 不, speakers of Mandarin in northern China often change the high tone of the numerals 七 *qī* 'seven' and 八 *bā* 'eight' into a rising tone before a following falling tone: we have not implemented this rule in the current version of the Mandarin synthesizer.

larger containing structures: the prosodic structures involved is related to the morpho-syntactic structure of the utterance by an algorithm described more fully in [Shih 1986]. For the purposes of the present discussion, note that the algorithm will follow the syntactic structure in grouping (polysyllabic) words together into a prosodic unit before it will group those words with other words into a larger prosodic unit. Thus in the utterance 小老鼠 *xiăo lăoshŭ* 'little rat', 老鼠 'rat' will be a unit that is contained inside the whole unit that contains 小老鼠 'little rat'; similarly 老鼠 will be a prosodic unit within the larger prosodic unit spanning the utterance 老鼠屎 *lăoshŭ shĭ* 'rat droppings':

(xiăo (lăoshŭ)) xiăo lao[2]shŭ 小老鼠 'little rat'
((lăoshŭ) shĭ) lao[2]shu[2] shĭ 老鼠屎 'rat droppings'

As one can see, this difference in structure leads to a difference in the application of third tone sandhi. In the case of 小老鼠, third tone sandhi applies first to the smallest unit *lăoshŭ*, changing the first syllable to *lao[2]*. In the next larger unit, the syllable *xiăo* no longer precedes a low tone, so third tone sandhi does not apply (is blocked). In 老鼠屎 *lăo* is changed to *lao[2]* in the same way, but here the larger structure is left-branching and the extra syllable is to the right. Thus the low tone of *shŭ* still precedes the low tone of *shĭ*, and so this syllable changes to *shu[2]*.

One commonly held misunderstanding about third tone sandhi is that the surface tonal pattern directly reflects the syntactic structure, and that the boundary strength between a sandhi/non-sandhi pair is always weaker than a non-sandhi/sandhi pair. The following examples show that this view is incorrect. In these two contrasting examples, the structures of the sentences are identical, but because of the tonal difference on the last syllable, the tonal outputs are quite different. Given the same subject/predicate boundary between *Lĭ* and *măi*, in one case tone sandhi applies to the subject *Li[2]*, while in another case tone sandhi applies to the verb *mai[2]*. This property of tone sandhi makes it difficult for statistical modeling, because the space of possible structures is large. Note that the effect is non-local in that the tone of final syllable ultimately determines the tone of the second syllable of the sentence. Longer sentences displaying exactly the same property could easily be constructed:

Lăo-Lĭ [$_{vp}$ măi hăo-jiŭ → Lao[2]-Li[2] măi hao[2]-jiŭ 老李買好酒 'Old Li buys good wine'
Lăo-Lĭ [$_{vp}$ măi hăo-shū → Lao[2]-Lĭ mai[2] hăo-shū 老李買好書 'Old Li buys good books'

Using a cyclic rule, we can easily derive the desired output. In the first sentence, *Lăo* and *hăo* assume the sandhi tone while the most deeply embedded structures *Lăo Lĭ* and *hăo jiŭ* are being scanned. The next level of scanning includes *măi hao[2] jiŭ*; since *hao[2]* now has a sandhi tone, this destroys the sandhi environment for *măi*. The final round of scanning includes the whole sentence, with the syllable pair of interest being *Lĭ măi*. Since the tone sandhi environment is met, the rule applies to produce *Li[2]*, despite the relatively strong boundary between subject and predicate. For the second sentence, *hăo* keeps its third tone because it is not followed by another third tone. Next, *mai[2]* assumes the sandhi tone since it precedes a third tone. Finally, when the syllable pair *Lĭ mai[2]* is being considered, tone sandhi is no longer applicable for *Lĭ*, since the tone sandhi environment was destroyed by the previous application to *mai[2]*.

The current implementation of third tone sandhi in the synthesizer does not build deep prosodic structures, but rather relies on the observation in [Shih 1986] that one possible pattern for Sandhi application is a first-pass application within prosodic words, and a second-pass phrasal application. This is implemented with the following two rules, which apply in the order written, the first of which is written so as to apply *only* within words, the second which applies across words. We presume a phonetic representation where tone symbols are written before the associated syllable:

$3 \rightarrow 2$ / ___ {Phoneme}* 3
$3 \rightarrow 2$ / ___ {Phoneme}* # 3
These rules are compiled into a transducer using the rewrite rule compiler.

Sandhi for *bù* 'not'. The special sandhi rule for 不 'not' is different from both third tone sandhi on the one hand, and the tonal variation for 一 'one' on the other, in that it makes reference to both lexical properties (i.e. the fact that the morpheme is 不 and not some homophonous morpheme such as 布 *bù* 'cloth'), and to phonological properties of the following syllable (i.e. that it has a falling tone). Such a rule could be implemented as a *two-level rule* [Koskenniemi 1983], but we have chosen the equivalent though formally different route of representing the 'underlying' pronunciation of 不 with a special tonal symbol '$2 \cup 4$', and then having a rule which expresses '$2 \cup 4$' as '2' before a following '4', and otherwise as '4'. Thus 不看 $bu_{2\cup4}$ *kàn* 'not look' becomes *bú kàn*.

The simple transducer that maps from characters to their basic pronunciations, is composed with the transducers implementing the tone sandhi rules just described to produce a transducer P, that maps from lexical representations into pronunciations. The entire transduction from text into phonemic representation can now be defined as $BestPath(BestPath(S \circ L \circ T) \circ P)$. For the input sentence 史立瑋有2隻老鼠。 this produces the phonetic transcription [3S% 4li 2xYeN # 2yO # 3lyaG # 1Z% # 2lW 3Su ɪ], using the phonetic transcription scheme introduced in Section 6.1. Once the best paths of the lexical and phonological lattices have been built, the information represented in those paths is placed into internal structures that are then passed to the next module of TTS, the module which builds prosodic phrases based on information about the location of those phrases as provided by the text-analysis module.

5.4 Future Work.

Two major deficiencies in the current text-analysis models for Mandarin are the prediction of prosodic phrase boundaries, and the treatment of homographs.

Prosodic phrase boundary prediction for the Bell Labs American English system, reported in [Wang and Hirschberg 1992], depends upon a Classification and Regression Tree (CART) model, trained on a tagged corpus, to predict the presence of prosodic phrase boundaries. The factors used in the CART model include information on the part of speech of words surrounding the putative boundary, as well as the distance of that boundary from the nearest punctuation mark. The model is able to predict phrase boundaries even when there is no punctuation to signal the presence of those boundaries, only whitespace between the adjacent words. An analogous solution could be used for Chinese. Indeed, as is shown in [Sproat and Riley 1996], it is possible to compile a CART tree into a weighted finite-state transducer, so the information incorporated in the decision tree could be used directly in the text-analysis system we have described.

As noted above, homographic characters are to some extent disambiguated using lexical information. However there are many cases where lexical constraints are lacking or at least too weak to be reliable. Consider the word 同行, which is *tónghánɡ* if it is a noun meaning 'colleague', but *tóngxínɡ* if it is a verb meaning 'travel together'. In principle, a reliable part-of-speech analyzer could help disambiguate these cases. So in a phrase like 在同行間的評價 *zài tónghánɡ jiān de*

píngjià (in colleague among 's evaluation) 'evaluation amongst colleagues', with the pronunciation *tónghán*g, the frame 在...間 'amongst' should in principle be a clear indicator that the contained phrase is a noun.[14] But the case is perhaps less clear for a phrase like 一路同行 *yílù tóngxíng* (one way together-travel) 'traveling together all the way', with the pronunciation *tóngxíng*: here 路 *lù* could in principle be either a noun or a measure word, and in either case purely grammatical constraints do not really suffice to rule out the nominal interpretation of 同行. In any event, it has been our experience with English that part-of-speech algorithms of the kind reported in [Church 1988] in general perform very badly at disambiguating homographs like *wound* (i.e., /waund/ as the past tense of *wind*; /wund/ as a noun). A partial solution to the problem for English has been discussed in [Yarowsky 1996]. 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. The method has been shown to have excellent performance in disambiguating English homographs, and some exploratory analysis using the method for Mandarin also seems very promising. In the Mandarin examples discussed above, 一路同行 is correctly classified as an instance of the *xíng* pronunciation due to the preceding 路 'way', which is statistically a reliable indicator of that pronunciation; similarly 在同行間的評價 is correctly assigned the pronunciation *háng*, because of the following character 間 'among'. As with the case of decision trees, it should be possible to represent the decision list as a WFST, in which case the results of training can be used directly in the text-analysis model discussed in this section.

6 Acoustic Inventory

Our synthesizer is an LPC-based concatenative system. The speaker upon which the synthesizer is based is a male speaker of Beijing Mandarin. Speech segments are stored in a table, and selected for synthesis at run time. These units are then concatenated together, with some speech signal manipulation to achieve the desired length, pitch, and amplitude. Most of the units are diphones, but a few triphones are included to avoid problematic diphones. As we have already noted in Section 3, the use of diphones rather than syllables distinguishes our system from most other systems for Mandarin.

6.1 Sound Inventory

Our current synthesizer distinguishes 43 phones for Mandarin, plus 8 extra ones to read English letters.[15] The Mandarin phones include 22 initial consonants, 3 glides, 3 consonantal endings, 11

[14] Still, this is not completely reliable since a deverbal noun could also appear in this position: 在行走間 *zài xíngzǒu jiān* 'while going'.

[15] Some English letter names have been incorporated into Chinese and can effectively be said to be part of the modern language: *X* 光 *eks-guāng* 'X-rays', *OK*, *PC*. In addition, a high percentage of Chinese text contains untransliterated

single vowels, and 4 diphthongs. These are listed in Table 2 together with 拼音 *pinyin* and 注音符號 *zhùyīn fúhào* (ZYFH) notations. Note that the correspondence between our system and *pinyin* or *ZYFH* is not one-to-one, due to the quasi-*phonemic* (rather than phonetic) nature of these transliteration schemes, and to phonetic ambiguities within the spelling systems. For the sake of clarity we present *pinyin* spellings in italics and present our internal representation in square brackets.

The consonant inventory is uncontroversial. Except for the addition of [v], which is used as a allophone of *w* in front of non-back vowels by northern speakers, our system is consistent with the traditional description of Mandarin sound inventory.[16] There are two vocalic endings *i* and *u*, which are not listed in our system because we treat diphthongs as whole units. Furthermore, Mandarin has three on-glides *y*, *w* and *yu*, which have similar formant target values to their corresponding high vowels *i*, *u* and *ü*.

The number of Mandarin vowels has been matter of heavy debate due to strong coarticulation effects between vowels and surrounding phonetic material. For example, most traditional descriptions consider the three variants of *a*, manifested in pinyin *ba*, *ban*, and *bian*, to be the same phoneme. Such a phonemically-oriented categorization is not suitable for a speech synthesizer, which must capture phonetic reality. So we separate the *a* sounds in *ba* and *ban* and represent them with different symbols [a] and [@]. At the same time, we merge the *a* in *bian* with the phone [e], since there is no acoustic difference between the two. Likewise, the letter *i* in pinyin represents three different sounds in *zi*, *zhi*, and *di*, which are also represented separately in our system.

The four diphthongs *ei*, *ai*, *ao*, and *ou* are represented as whole units in our system.

6.2 Diphone Units

A complete speech synthesizer must be able to handle all (linguistically possible) phone-to-phone transitions. It is easy to calculate the required diphones in Mandarin, because the inventory of syllables is rather uncontroversial (at least at the phonemic level): the same thing cannot be said, for example, about English. One requires all syllable internal units, including consonant–glide, consonant–vowel, glide–vowel, and vowel–ending; plus cross-syllable units, which include the transitions between all units that can end a syllable (vowel, ending, or silence) and all units that can begin a syllable (silence, consonant, glide, or vowel). The number of required diphones can be drastically reduced by combining units that are acoustically indistinguishable. For example, given two nasal endings [N] and [G] (*n* and *ng*), and ten stops/affricates (*p*, *t*, *k*, *b*, *d*, *g*, *z*, *c*, *zh*, *ch*), we in theory will need twenty diphonic units. But in fact, all Mandarin stops/affricates are voiceless and begin with a section of silent closure, so we can simply collect two units — [N–*] [G–*], and use them for all of the stops and affricates. One can go further in eliminating vowel–stop and vowel–affricate units if the size of the inventory is an issue; this was what was done in the 1987 Bell Labs Mandarin system. However, this does result in a slight loss of quality since there is of course some coarticulation between the vowel and the following consonant: we have therefore kept these

foreign (usually English) words. Our text-analysis module actually contains a dictionary giving pronunciations of about twenty thousand frequent English words, but we currently do not have the acoustic units necessary to handle such words elegantly within the system.

[16] Northern Mandarin also has a retroflex 兒 *-r* ending, resulting from suffixation, in addition to the nasal endings *n* and *ng*.

	Bell Labs System	Pinyin	ZYFH
Consonant	p	p	ㄆ
	t	t	ㄊ
	k	k	ㄎ
	b	b	ㄅ
	d	d	ㄉ
	g	g	ㄍ
	z	z	ㄗ
	j	j	ㄐ
	Z	zh	ㄓ
	c	c	ㄘ
	C	ch	ㄔ
	q	q	ㄑ
	f	f	ㄈ
	s	s	ㄙ
	x	x	ㄒ
	S	sh	ㄕ
	h	h	ㄏ
	n	n	ㄋ
	m	m	ㄇ
	l	l	ㄌ
	r	r	ㄖ
	v	w(ei)	ㄨ
Ending	N	(i)n	ㄣ
	G	ng	ㄥ
	L	(ya)r	ㄦ
Glide	y	y, (x)i(e)	ㄧ
	w	w, (sh)u(a)	ㄨ
	Y	yu, (j)u(an)	ㄩ
Vowel	i	i	ㄧ
	$	(c)i	ㄗ, ㄘ, ㄙ
	%	(sh)i	ㄓ, ㄔ, ㄕ
	u	u	ㄨ
	U	ü	ㄩ
	e	(xi)e	ㄝ
	E	(k)e, (k)en, (k)e(ng)	ㄜ
	o	(du)o	ㄛ
	R	er	ㄦ
	a	a	ㄚ
	O	ou	ㄡ
	A	ei	ㄟ
	I	ai	ㄞ
	W	ao	ㄠ
	@	a(n)	ㄢ

Table 2: Phonemes

units in the current version.

Another time and space saving strategy we adopted for the 1987 version was to eliminate glides from the sound inventory, using the corresponding high vowels in their place. About eighty diphones can be eliminated that way. We decided to revive the glides for the current version, again to improve quality at a cost of a larger inventory. Interestingly, the problem in combining glides and vowels is not in formant incompatibility. The major problem lies in the difficulty in modeling cross-syllable units with within-syllable unit, and vice versa. For example, the vowel to vowel unit [u–A] is typically pronounced with a glottal stop at the syllable boundary. If such a unit is used for the glide to vowel unit [w–A] for the syllable *wei*, the presence of the glottal stop will be noticeable. So care must be taken to collect units without intervening glottal stops. But if one does that, then one has the opposite problem in that some syllable-boundaries may be hard to detect in the synthetic output: as a result, sequences like 一五一十 *yīwǔ yīshí* (one five one ten) 'systematically' were hard to understand in the 1987 version since the first three syllables were blurred together. The 1987 version contained 492 diphones (note that this number is only marginally higher than the 410 syllable count). The current Bell Labs Mandarin system contains 960 diphones plus a few triphones. The 960 units include diphones needed to read English letters.

6.3 Diphone-set Expansion and Triphones

A diphonic system works well when the edges of concatenating units are similar. Moreover, in the case of vowels and glides, the formant trajectories should match as well. This ideal situation may not materialize for two reasons. First, there is a lot of variation in speech production in that two recordings of the same phone in the same context may nonetheless have enough differences to cause problems in diphone concatenation: thus two recordings by the same speaker of the word *cat* may be sufficiently different that there will be a noticeable discontinuity if one concatenates the [k–a] unit from the first utterance with the [a–t] unit of the second. This problem can be minimized by making a large number of recordings: usually, with a sufficient number of utterance one is able to find at least one unit that matches reasonably well with other units with which it must be concatenated.

The second problem is systematic coarticulation that introduces strong coloring to some diphone units, making them poor concatenative units with other *normal* units. This problem cannot be solved by simply making further recordings, since the problem relates to fundamental phonetic properties of the language, and not accidental differences in pronunciation. Even if a speaker may occasionally produce the unit without the coarticulation effect so that a particular diphone falls into the acceptable range for concatenation, such units may not be accepted as a natural rendition by listeners. Two solutions to this problem are, firstly to increase the set of phones so as to arrive at a better classification of the phonetic variation, and secondly to construct triphonic units. We will see instances of both these solutions in the treatment of units containing the central vowel.

There are two problematic areas in Mandarin for diphone collection [Shih 1995]. The first of these involves glides. Glides are short with fast-moving formant trajectories, their formant values being heavily influenced by the following vowel. A consonant-glide unit such as [d–w] taken from the [o] context may not connect well with [w–A]. One easy fix is to collect triphones such as [d–w–o] and [d–w–A]. However, the connectivity problem here is less severe than the problem with [w–E–N] to be discussed below, so we have chosen not to implement these sequences as triphones in the

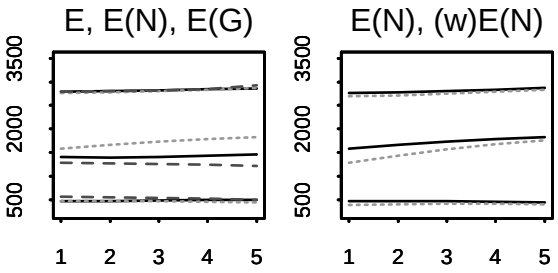

Figure 4: Coarticulation effects on [E]

present version.

The second problem involves the central vowel *e* as in 哥 *gē* 'brother', which we represent as [E]. It is not surprising that the central vowel, or the schwa, is a problematic sound in English, since schwa in English is always unstressed. It is interesting that the central vowel presents a problem to Mandarin as well, since in Mandarin the central vowel [E] can be fully stressed. Still, even stressed central vowels show a considerable amount of coarticulation with preceding glides and with following nasal endings. Figure 4 shows the major coarticulation effects on [E]. Trajectories of the first three formants of [E] in different contexts are compared. The five points on the x-axis correspond to the 20%, 35%, 50%, 65%, and 80% points of the vowel duration. Averaged formant frequencies are plotted on the y-axis. In the left-hand panel, solid lines are the average values of F1, F2, and F3 of the all the [E]'s in our database. Dotted line are the average formant values of all the [E]'s in the [EN] context (central vowel followed by an alveolar nasal coda), and dashed lines are the average formant values of all the [E]'s in the [EG] context (central vowel followed by a velar nasal coda). The F2 of [E] in the [EN] context is significantly higher than that of the average [E], while the F2 of [E] in the [EG] context is significantly lower than that of the average [E]. The directions of the coarticulation effects are expected and are perfectly consistent with the anticipated tongue positions of the following sound. However, the coarticulation effects start early, already present 20% into the vowel. Furthermore, the magnitude of the effects are strong. The difference in the F2 of [E] in [EN] and [EG] ranges from 300 Hz at the 20% point to 600 Hz at the 80% point. If such predictable discrepancies of the same phonemic sound are not taken care of it can severely hamper the quality of a concatenative speech synthesizer. Take [pEN] *pen* as an example. It is synthesized by connecting two units [p–E] and [E–N]. If the unit [p–E] is taken from the syllable [pEG], its F2 will be too low for [E–N]. On the other hand, if the unit [p–E] is taken from the [EN] context, then its F2 will be too high, and its rising F2 pattern incompatible with [E–G]. To ensure smooth connection, [E] in the [EN] context should be considered a separate sound from the rest of the [E] samples.

The right-hand panel of Figure 4 shows additional coarticulation effects caused by the back rounded onglide [w]. The solid lines plots the trajectories of the first three formants of all the [E]'s in the [EN] context, in contrast to the formant trajectories of [E]'s in the [wEN] context, which are ploted in dotted lines. Note that a preceding [w] effectively lowers F2 for the first 80% of the vowel, resulting in a drastic rising slope of F2. Fast moving formant trajectories like this are incompatible with a steady state vowel, making [wEN] a poor source for a generic diphone unit [E–N], and making [w–E–N] a good candidate for a triphonic entry. Indeed, our study shows that

[w–E–N] is the only triphone that is absolutely needed for Mandarin.[17]

Thus, based on the coarticulation effects just discussed, we collected a triphone unit [w–E–N], and two sets of diphone units for [E], one set being taken from and used in the [N] context, the other set being taken from and used in all other contexts.

7 Duration

The duration module assigns a duration value to each phoneme. The value is estimated from various contextual factors by multiplicative models that are fitted from a speech database collected specifically for the study of duration. The general design of the model and research methodology were reported in [van Santen 1994], and the Mandarin study was reported in more detail in [Shih and Ao 1996]. As we explained in Section 3, the difficulty in prosodic modeling lies in the large number of factor combinations. We approach this problem from both ends, maximizing the coverage in our database on the one hand, and estimating unseen data points on the other hand.

The first stage taken in producing a duration model involves constructing a set of sentences to be read by the speaker, which cover all the desired factor combinations. This preparation involves the following steps:

1. Automatically transcribe short paragraph-sized sentences in the on-line text database into broad phonetic transcription, using the text-analysis model described in Section 5.[18]

2. Code each segment with relevant factors, including the identity of the segment and its tone, the category of the previous segment and tone, the category of the next segment and tone, the syllable type, and the distance from the beginning and end of the word, phrase, and utterance. We then combine the coded factors into factor triplets, each containing the identity of the current segment, its tone, and another factor.

3. Feed the coded factor triplets to a *greedy algorithm* [Cormen *et al.* 1990]. In each round, a sentence with the largest number of unseen factor triplets is chosen, until all factor triplets are covered, or the number of sentences has reached a preset limit. In our case, a total of 424 sentences consisting of 48706 segments covering all the factor triplets that are present in the input, were chosen from a pool of 15620 sentences.

4. Record the 424 sentences by the same male Beijing speaker used for the concatenative unit inventory.

5. In some cases the speaker uttered the sentence in a different way from what was expected from the automatic transcription. This required some recoding of the transcriptions in order to reflect what was actually said. Also added was information on prosodic phrasing and three levels of prominence, neither of which was generally predictable via the automatic transcription.

[17] Our system contains a number of other triphones, but those are due to accidental bad connection of diphones, rather than consistent, unavoidable coarticulation effects.

[18] The online corpus used for duration modeling was the ROCLING newspaper corpus.

o	%	E	$	i	U	u	e
99	109	113	116	120	121	121	128
R	A	O	I	F	W	a	
134	135	138	147	149	155	160	

Table 3: Intrinsic Duration of Vowels in Msec

h	f	S	x	s
98	100	113	119	122

Table 4: Intrinsic Duration of Fricatives in Msec

This procedure resulted in a data matrix including every phone in the database, its coded contextual factors and, of course, its duration. Preliminary exploratory data analysis was performed on this data matrix and data points were separated into classes based on this analysis so as to minimize factor interaction within class. For example, we know that final lengthening affects onset and coda consonants in different ways, so onset and coda consonants are separated into different classes for modeling. The eventual classes that were decided upon were: vowel, glide, closure of stop and affricate, burst and aspiration of stop and affricate, fricative, sonorant consonant, and coda. Multiplicative models are fitted for each class separately to predict the duration.

We compare near minimal pairs in the database and interpolate the values for missing cells. Below is a much-simplified example of the estimation of missing values, namely a stressed [S] in a certain segmental, tonal, and prosodic environment. If the durations of unstressed [f], [S], and [s] in the same environment are 100, 110, and 120 msec respectively, we hypothesize that the relative scale of these three sounds is in the ratio $1 : 1.1 : 1.2$ and that the scale should hold under the influence of stress as well. If the stressed [f] and stressed [s] in the same contextual environment are 120 and 144 msec respectively, we deduce that stress in that environment has the effect of increasing the duration value by 20%, so the value of the missing stressed [S] should be 132 msec. Our data show very clear patterns of such *single factor independence*, and this is an important theoretical assumption in estimating duration values of missing data points.

Among the fourteen factors that we investigated, the identity of the following tone is the only one that shows nearly no effect on any classes of sound. The factors that consistently have a strong effect on all classes of sounds are the identity of current segment and prominence. All the other factors have some effect on some classes but not on others.

The intrinsic duration (IDur) of vowels, fricatives, bursts/aspirations, and closures estimated from our database are summarized in Table 3, Table 4, Table 5, and Table 6 respectively.

At runtime, segment identities and relevant factors are computed and factor coefficients multiplied to derive the estimated duration. For example, if a sentence contain a single sound [i4], its duration will be computed as follows:

b	d	g	Z	z	j	p	t	k	C	c	q
11	13	21	29	43	46	80	80	86	95	99	113

Table 5: Intrinsic Duration of Burst-Aspiration in Msec

Ccl	ccl	qcl	zcl	tcl	jcl	Zcl	kcl	dcl	gcl	pcl	bcl
13	13	15	15	16	16	17	18	18	19	20	21

Table 6: Intrinsic Duration of Closure in Msec

$$
\begin{aligned}
Dur_i(229.57msec) = \\
IDur_i(120.08msec) \\
\times F2_{tone[4]}(1.057) \times F3_{previous-phone[null]}(1.24) \\
\times F4_{previous-tone[null]}(0.99) \times F5_{next-phone[null]}(1.1) \\
\times F6_{next-tone[null]}(1.03) \times F7_{stress[1]}(1.09) \\
\times F8_{preceding-syllable-in-word[0]}(1.01) \\
\times F9_{following-syllable-in-word[0]}(1.03) \\
\times F10_{preceding-syllable-in-phrase[0]}(1.02) \\
\times F11_{following-syllable-in-phrase[0]}(1.2) \\
\times F12_{preceding-syllable-in-utterance[0]}(0.96) \\
\times F13_{following-syllable-in-utterance[0]}(0.8) \\
\times F14_{syll-type[v]}(1.22)
\end{aligned}
$$

8 Intonation

The intonation module computes F0 values for the synthetic speech. Since Mandarin is a tone language, one important task is to capture tones and tonal variations.

We start out with a basic representation of Mandarin tones, derived from text analysis, which are then passed to a set of tonal coarticulation rules that take care of changes that result from the mechanics of tonal concatenation. At this stage, we derive a sequence of tonal targets that is devoid of sentence intonation and emphatic expression, but otherwise represents smoothly connected tones. Effects of sentence intonation and emphasis can then be added. We have implemented declination and downstep, two major factors that causes pitch to drop over the course of an utterance. We plan to add phrasing and emphasis effects in the near future.

Mandarin has four tones, plus the possibility that a syllable is reduced and carries no lexical tones, a phenomenon traditionally referred to as *neutral tone*, or *tone 0*. Table 7 lists the tone names, their description, and the tonal target we assign to the tones. The letters H, M, L represent

Tone Symbol	Tone 1	Tone 2	Tone 3	Tone 4	Tone 0
Tone Name	High Level	Rising	Falling-rising	Falling	Neutral
Targets	H H	M H	M L (M+)	H+ L	M

Table 7: Mandarin Lexical Tones

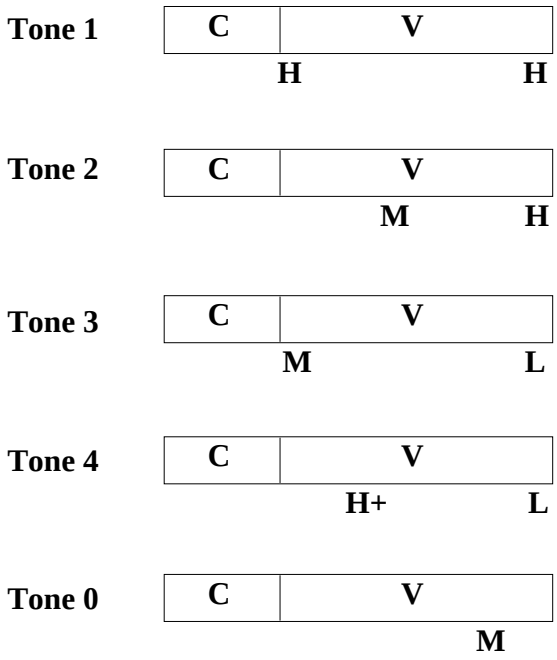

Figure 5: Alignment of Tonal Targets

high, mid and low targets respectively. Figure 5 shows the schematic alignment of targets to syllables.

The phonetic realization of the four lexical tones are relatively stable, and is adequately reflected in their names. The first target of tone 2 is placed near the center of the rime to free the first portion for coarticulation with the preceding tone. The first target of tone 4 is higher than the normal H level, hence the notation *H+*. It is placed somewhat into the rime to create the peak associated with tone 4.

The phonetic shape of the neutral tone is unstable, and it primarily depends on the tonal identity of the preceding tone. In general, neutral tones following tones 1, 2, and 4 have a falling shape but neutral tones following tone 3 have a rising shape. For the four lexical tones, we assign multiple tonal targets that span the duration of the rime, which ensures the stability of the tone shapes. The single M target of tone 0 is associated to a point near the end of the rime. The single target captures the unstable tone shape, its association near the end of the rime accounting for the fact that the shape is largely determined by the preceding tone. When the final target of the preceding tone is L, as in the case of tone 3, the pitch rises toward M, but when the final target of the preceding tone is H, as in the case of tone 1 and tone 2, the pitch falls to M. A neutral tone

following a tone 4 requires special treatment. A non-final tone 4 differs from a final one in that it ends in a M target, rather than L. Given that, the predicted shape of a neutral tone following tone 4 is level, not the observed falling. This will be handled by the coarticulation rules.

Tones in connected speech deviate from their canonical forms considerably. The changes may come from tonal coarticulation [Shih 1987, Shen 1990b, Xu 1993] or phrasal intonation [Shen 1985, Shih 1987, Shih 1988, Gårding 1987], as well as various lowering effects. Our rules are primarily based on our own instrumental studies. Both tonal target and target alignment may be changed by tonal coarticulation rules, but sentence level scaling is accomplished by changing the pitch height of H, M and L targets.

The most important tonal coarticulation rules are listed below with brief descriptions.

1. *If tone 4 is not sentence-final, change its L target to M.*

   Our phonetic data shows that non-final tone 4 consistently falls to M, rather than to L. So the majority of tone 4 actually have the tonal targets HM. This rule replaces a more limited rule that changes a tone 4 only before another tone 4 used in many systems [Zhang 1986, Shen 1990b, Lee *et al.* 1989, Chiou *et al.* 1991, Hsieh *et al.* 1989], apparently following the early report in [Chao 1968]. Listening tests show that restricting the rule to the tone 4 context results in over-articulation of tone 4 in other tonal contexts, and the result of our more general version of the rule is more natural.

2. *Given a tone 2, if the following tone starts with a H target, then delete the H of tone 2.*

   Comparing tone 2 before a H target and tone 2 before a L target, the rising slope before a H target is more gradual, and the pitch does not reach the expected H at the end of the tone 2 syllable. This is not a simple dissimilation effect. The desired pitch contour can be generated easily by assuming that the H target of tone 2 is absent.

3. *If a tone 0 follows a tone 4, change its M target to L.*

   Tone 0 should have a low falling pitch when it follows a tone 4. Since non-final tone 4 ends in M, assigning a L target on the tone 0 syllable will generate the desired falling contour.

4. *If the following tone start with a H target, then the L target of tone 3 is raised:* $L = L + 0.15(M - L)$

   The L target of tone 3 is raised when the following target is H. This is an assimilation effect.

5. *If tone 3 is sentence-final, add a M target, assign its value to be* $M = M + 0.3(H - M)$

   It is a well-known phenomenon that tone 3 in the sentence-final position or in isolation has a falling-rising shape, while other tone 3 have a low-falling shape, hence the name 'half tone 3' [Chao 1968]. This rule adds the rising tail in the sentence-final condition. The pitch rises above the M point, but not as high as H.

Coarticulation rules that are often found in other systems [Hsieh *et al.* 1989, Lee *et al.* 1989] but are excluded here are sandhi rules, which are handled in the text-analysis component, and various lowering rules, which we attribute to two separate effects of declination and downstep.

Downstep [Liberman and Pierrehumbert 1984] is a lowering effect triggered by L (or non-high) targets. For example, in a combination H L H, the second H is lower than the first H. Declination in the early literature [Lieberman 1967, Ö hman 1967] refers to the general downtrend of pitch over the course of an utterance, including downstep. Following [Liberman and Pierrehumbert 1984], we modify the definition of declination as the lowering effect after the downstep effect has been factored out. Now, Liberman and Pierrehumbert argued strongly that once downstep and other effects like final lowering were factored out, there is in fact no evidence for a residual declination factor. Contra this position, we find considerable evidence of declination in Chinese. In sentences totally lacking non-high targets, namely, sentences consisting of tone 1 syllables, there is still a gradual lowering of pitch.

Our default setting of declination is to lower the pitch of subsequent syllables by 1% of the distance between H and M. We implement downstep as an exponential decay that asymptotes to $(M + a)$, where a is a constant that we set initially at 10 Hz:

$$H = H - d_i(H - M + a)$$

The symbol d denotes the downstep factor that varies with tonal target identity i. Our investigation shows that L target causes more downstep on the following tone than M. Tone 3 with a L target causes the most amount of downstep, tone 2 and 4 with M value cause less downstep. Tone 1, lacking a L or M target, causes no downstep.

Emphasis causes expansion of pitch range, and compression of pitch range after the emphasis. The effect is most notable in the raising of H targets. One difference in our findings from previous reports [Zhang 1986, Gårding 1987] is in the scope of the emphasis effect. We find that emphasis is not strictly a local effect manifested only on the emphasized syllable. Pitch raising due to emphasis starts from the emphasized syllable, but the duration of the effect depends on tonal composition. Pitch returns to normal when it hits the first L or M target, which includes the M or L target of tone 4 if this tone 4 is emphasized. If there is a long sequence of H targets following the elevated pitch of the emphasized syllable, the unemphasized H targets remain H for a while and drift down in a similar fashion as declination, but at a faster rate. The emphasis effect does not cross phrase boundaries, so if a speaker pauses after the emphasized word, the pitch will return to normal immediately.

This separation of tonal coarticulation factors from intonational factors reflects the way we believe intonation works in a tone language. On the surface, there are in principle an infinite variety of tonal values. But underlyingly there are clearly only five tonal categories. Each of these is related to a tonal template which captures the invariant aspects of the tone. The tonal coarticulation rules ensure that tonal connections are smooth, but at this stage pitch values are still represented in relative terms as shades of H, M, and L. We then assign values to the targets, and adjust their values according to the intonation effects that are suitable for the context, such as question intonation, declarative intonation, emphasis, and paragraph modeling, with appropriate lowering or raising effects. Although the broad characteristics of many of these intonation effects are known [Chao 1933, Shih 1988, Shen 1990a, Tseng 1981], the level of details and control that are required for a text-to-speech system remain to be worked out. Our design makes it easy to add layers of intonational effects when the research results become available.

9 Conclusions

As we believe this paper has made clear, there has been substantial progress on Mandarin text-to-speech synthesis over the last two decades, starting from the mid seventies, to the present. We have progressed from systems which could transform annotated phonetic transcriptions into barely intelligible speech, to systems which can take Chinese written in normal Chinese orthography and transform it into speech that is highly intelligible, though certainly still mechanical in quality.

Where do we go from here? To a large extent the problems that remain to be solved for Mandarin are the same as the problems that are faced by synthesizers for any language. Text analysis capabilities are still rudimentary, with particular difficulties arising with the prediction of prosodic phrase boundaries, and the selection of appropriate levels of prominence for words.

The selection of appropriate intonational contours is also a difficult problem. It is interesting to note, though, that Mandarin is to some extent easier in this regard than, say, English. In English, essentially none of the F0 variation in a sentence is lexically determined, all of it being determined by the selection of pragmatically appropriate accent types. In Mandarin, on the other hand, a great deal of the F0 variation *is* lexically determined, with phrase-level intonation being responsible not so much for the overall shape of the F0 pattern, as for the relative values of the lexically determined peaks and valleys. This is not to marginalize the importance of work on Mandarin intonation, but it does mean that it is easier in Mandarin than in English to produce a reasonable-sounding F0 contour.

A third major area for improvement is the quality of the actual speech output. Concatenative systems continue to be plagued by subtle problems relating to slight incompatibilities in concatenated units, and further research is needed on improved methods for selecting and cutting units. There is also a ripe area of research in what are sometimes termed 'hybrid' synthesis methods, whereby the formant structure of adjacent units can be manipulated during the concatenation process, so as to produce smoother transitions, or even to implement phonetic reductions or coarticulation: for example, with adequate abilities to modify units 'on the fly' one could possibly avoid the increase in the number of [E] units discussed in Section 6.3. The synthesis methods themselves require further work. Waveform-based approaches are showing more and more promise and may eventually become more desirable than LPC-based methods. Finally, the best synthesis method in the world will not result in natural-sounding speech if adequate *linguistic* controls of various aspects of voice quality are not available. To take just one example, it was shown in [Pierrehumbert and Talkin 1992] that intervocalic *h* in English can trigger a variety of glottal states ranging from a full aspiration, to slight glottalization or even complete lenition, depending upon the prosodic conditions. It certainly seems to be the case that no commercially available TTS systems implement this variation.

To quality of output, one can also add variety as a goal. Most concatenative systems are based on one or at most two voices: in the latter case the voices are usually male and female, as in the case of the Bell Labs American English system. By modifying such parameters as the pitch range, the vocal tract length, and the speech rate, it is possible to some extent to modify these voices so that they sound like different people, but the results of these modifications are usually more comical than realistic. Two approaches to improving this situation suggest themselves. The first is to provide a set of tools that makes it virtually painless to record a new voice and build an inventory for that voice. Systems for automatically cutting diphones from recorded speech given

an *orthographic* transcription of that speech have been reported,[19] but we have not seen evidence that these approaches are yet capable of producing an inventory that has the same quality as one that involves at least some handwork. Nonetheless, we expect matters to improve in this area in the next few years. The second is to investigate whether one can design transformations that are capable of mapping an inventory of one speaker onto an inventory that sounds plausibly like another speaker — either an actual person or a fictitious one.

We expect that synthesizers will continue to improve in all of these areas, though it will, in our view, be many years before TTS systems become indistinguishable from humans, and thus pass the Turing test for speech synthesis. One thing that is certain is that speech synthesis will rapidly become a technology that is familiar not only to a select group of researchers, computer afficionados, and people with special disabilities, but indeed to the public at large. The applications of TTS are numerous. Systems for English have been used as aids for the handicapped, including readers for the blind, and speech output devices for people who have lost their ability to speak (the Cambridge University physicist Stephen Hawking being the most famous example of this); they have been used as email readers allowing one to access one's email over the telephone; they have been used in so-called 'reverse-directory' applications, where a caller uses a telephone touch-tone pad to key in a telephone number or a postal code, and the system reads back the name and address of the entry, or entries, that match the number; there is interest in using TTS systems for reading directions as part of navigational systems for drivers; and they have been used in such futuristic applications as speech-to-speech translation.[20] Mandarin TTS systems have been applied in some of these domains, and their use in various applications will surely increase over the coming decade. Undoubtedly there will also be a demand for synthesizers in other Chinese languages. As we have mentioned there has been some work on Taiwanese and Cantonese, and we certainly expect that work on regional languages will continue.

10 Acknowledgments

We acknowledge BDC (Taiwan) for the 'Behavior Chinese-English Electronic Dictionary' used in our word segmenter, and ROCLING for providing us with the text database used in our duration study. J. S. Chang kindly provided us with his database of personal names. We also wish to thank Rob French for manually segmenting our acoustic inventory database, and Jan van Santen for duration analysis tools, as well as general advice on duration-related issues. Benjamin Ao also contributed substantially to the work on duration modeling. Mike Tanenblatt was responsible for producing robust versions of the multilingual intonation and duration modules. Finally, we extend our thanks to Shengli Feng for providing the voice for our synthesizer.

REFERENCES

Jonathan Allen, M. Sharon Hunnicutt, and Dennis H. Klatt. *From Text to Speech: The MITtalk*

[19] For example, Silverman discussed the Apple work in this area at the 1994 ESCA/IEEE Workshop on Speech Synthesis in New Paltz, New York.

[20] Our Mandarin and English systems are being used in a limited-domain speech translation system under development at AT&T Research.

System. Cambridge University Press, Cambridge, UK, 1987.

Benjamin Ao, Chilin Shih, and Richard Sproat. A Corpus-Based Mandarin Text-to-Speech Synthesizer. In *Proceedings of the International Conference on Spoken Language Processing*, (Yokohama, Japan), pp. 1771–1774, ICSLP, 1994.

Leonard Bloomfield. *Language*. Holt, Rinehart and Winston, New York, 1933.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth & Brooks, Pacific Grove CA, 1984.

Lianhong Cai, Hao Liu, and Qiaofeng Zhou. Design and Achievement of a Chinese Text-to-Speech System under Windows. *Microcomputer*, Vol. 3, 1995.

Ngor-Chi Chan and Chorkin Chan. Prosodic Rules for Connected Mandarin Synthesis. *Journal of Information Science and Engineering*, Vol. 8, pp. 261–281, 1992.

Yueh-Chin Chang, Yi-Fan Lee, Bang-Er Shia, and Hsiao-Chuan Wang. Statistical Models for the Chinese Text-to-Speech System. In *Proceedings of the 2nd European Conference on Speech Communication and Technology*, pp. 337–340, 1991.

Yueh-Chin Chang, Bang-Er Shia, Yi-Fan Lee, and Hsiao-Chuan Wang. A Study on the Prosodic Rules for Chinese Speech Synthesis. In *International Conference on Computer Processing of Chinese and Oriental Languages*, pp. 210–215, 1991.

Jyun-Sheng Chang, Shun-De Chen, Ying Zheng, Xian-Zhong Liu, and Shu-Jin Ke. Large-Corpus-Based Methods for Chinese Personal Name Recognition. *Journal of Chinese Information Processing*, Vol. 6, No. 3, pp. 7–15, 1992.

Yuen Ren Chao. Tone and Intonation in Chinese. *Bulletin of the Institute of History and Philology, Academia Sinica*, Vol. 4, No. 2, pp. 121–134, 1933.

Yuen Ren Chao. *A Grammar of Spoken Chinese*. University of California Press, 1968.

F. Charpentier and E. Moulines. Pitch-synchronous Waverform Processing Techniques for Text-to-Speech Synthesis Using Diphones. *Speech Communication*, Vol. 9, No. 5/6, pp. 453–467, 1990.

Sin-Horng Chen and Yih Ru Wang. Vector Quantization of Pitch Information in Mandarin Speech. *IEEE Transactions on Communications*, Vol. 38, pp. 1317–1320, 1990.

Sin-Horng Chen, Saga Chang, and Su-Min. Lee. A Statistical Model Based Fundamental Frequency Synthesizer for Mandarin Speech. *Journal of the Acoustical Society of America*, Vol. 92, No. 1, pp. 114–120, 1992.

Sin-Horng Chen, Shaw Hwa Hwang, and Chun-Yu Tsai. A First Study of Neural Net Based Generation of Prosodic and Spectral Information for Mandarin Text-to-Speech. In *Proceedings of IEEE ICASSP*, Vol. 2, pp. 45–48, 1992.

Sin-Horng Chen, Shaw-Hwa Hwang, and Yih-Ru Wang. An RNN-based Prosody Information Synthesizer for Mandarin Text-to-Speech. *IEEE Audio and Speech Processing*, forthcoming.

Hong-Bin Chiou, Hsiao-Chuan Wang, and Yueh-Chin Chang. Synthesis of Mandarin Speech Based on Hybrid Concatenation. *Computer Processing of Chinese and Oriental Languages*, Vol. 5, No. 3/4, pp. 217–231, 1991.

John Choi, Hsiao-Wuen Hon, Jean-Luc Lebrun, Sun-Pin Lee, Gareth Loudon, Viet-Hoang Phan, and Yogananthan S. Yanhui, A Software Based High Performance Mandarin Text-to-Speech System. In *Proceedings of ROCLING VII*, pp. 35–50, 1994.

Noam Chomsky and Morris Halle. *The Sound Pattern of English.* Harper and Row, New York, 1968.

Min Chu and Shinan Lü. High Intelligibility and Naturalness Chinese TTS System and Prosodic Rules. In *Proceedings of the XIII International Congress of Phonetic Sciences*, (Stockholm), pp. 334–337, 1995.

Kenneth Church. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, (Morristown), pp. 136–143, Association for Computational Linguistics, 1988.

Cecil Coker. Speech Synthesis with a Parametric Articulatory Model. In *Speech Symposium*, (Kyoto), 1968. Reprinted in Flanagan, James and Rabiner, Lawrence (eds.) *Speech Synthesis*, Dowden, Hutchinson and Ross, Stroudsburg, PA, 135–139, 1973.

T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms.* The MIT Press, Cambridge, Massachussetts, 1990.

John DeFrancis. *The Chinese Language.* University of Hawaii Press, Honolulu, 1984.

N. R. Dixon and H. D. Maxey. Terminal Analog Synthesis of Continuous Speech Using the Diphone Method of Segment Assembly. *IEEE Transactions on Audio Electroacoustics*, Vol. AU-16, pp. 40–50, 1968.

Gunnar Fant. *Acoustic Theory of Speech Production.* Mouton, The Hague, 1960.

Long Feng. Beijinghua Yuliu Zhong Sheng Yun Diao de Shichang (Duration of Consonants, Vowels, and Tones in Beijing Mandarin Speech). In *Beijinghua Yuyin Shiyanlu (Acoustic Studies of Beijing Mandarin)*, pp. 131–195, Beijing University Press, 1985.

Eva Gårding. Speech Act and Tonal Pattern in Standard Chinese: Constancy and Variation. *Phonetica*, Vol. 44, pp. 13–29, 1987.

W. L. Henke. Preliminaries to Speech Synthesis Based upon an Articulatory Model. In *Proceedings of the IEEE Conference on Speech Communication Process*, pp. 170–182, IEEE, 1967.

Ching-Jiang Hsieh, Chiu-Yu Tseng, and Lin-Shan Lee. An Improved Chinese Text-to-Speech System Based on Formant Synthesis Techniques. *Computer Processing of Chinese and Oriental Languages*, Vol. 4, No. 2/3, pp. 153–170, 1989.

Tai-Yi Huang, Cai-Fei Wang, and Yoh-Han Pao. A Chinese Text-to-Speech Synthesis System Based on an Initial-Final Model. *Computer Processing of Chinese and Oriental Languages*, Vol. 1, No. 1, pp. 59–70, 1983.

James Hurford. *The Linguistic Theory of Numerals.* Cambridge University Press, Cambridge, 1975.

Shaw Hwa Hwang and Sin-Horng Chen. Neural Network Synthesiser of Pause Duration for Mandarin Text-to-Speech. *Electronic Letters*, Vol. 28, No. 8, pp. 720–721, 1992.

Shaw Hwa Hwang and Sin-Horng Chen. Neural-network-based F0 Text-to-Speech Synthesiser for Mandarin. *IEE proceedings: Vision, Image and Signal Processing*, Vol. 141, No. 6, pp. 384–390, 1994.

Shaw Hwa Hwang and Sin-Horng Chen. A Prosodic Model of Mandarin Speech and its Application to Pitch Level Generation for Text-to-Speech. In *Proceedings of IEEE ICASSP*, Vol. 1, pp. 616–619, 1995.

Shaw-Hwa Hwang, Yih-Ru Wang, and Sin-Horng Chen. A Mandarin Text-to-Speech System. In *Proceedings of ICSLP 96*, 1996.

Hector Javkin, Kazue Hata, Lucio Mendes, Steven Pearson, Hisayo Ikuta, Abigail Kaun, Gregory DeHaan, Alan Jackson, Beatrix Zimmermann, Tracy Wise, Caroline Henton, Merrilyn Gow, Kenji Matsui, Nariyo Hara, Masaaki Kitano, Der-Hwa Lin, and Chun-Hong Lin. A Multilingual Text-to-Speech System. In *Proceedings of ICASSP*, pp. 242–245, 1989.

Gwo-Hwa Ju, Wern-Jun Wang, Chi-Shi Liu, and Wen-Hsing Lai. Yi Tao Yi Duomaichong Jifa Yuyin Bianmaqi Wei Jiagou zhi Jishi Zhongwen Wenju Fan Yuyin Xitong (A Real-time Implementation of Chinese Text-to-Speech System Based on the Multi-pulse Excited Speech Coder). *(TL Technical Journal)*, Vol. 21, No. 4, pp. 431–440, 1991.

Ronald Kaplan and Martin Kay. Regular Models of Phonological Rule Systems. *Computational Linguistics*, Vol. 20, pp. 331–378, 1994.

J. Kelly and L. Gerstman. An Artificial Talker Driven from Phonetic Input. *Journal of the Acoustical Society of America*, Vol. Supplement 1 33, p. S35, 1961.

Dennis Klatt. Review of Text-to-Speech Conversion for English. *Journal of the Acoustical Society of America*, Vol. 82, pp. 737–793, 1987.

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

R. Kurzweil. The Kurzweil Reading Machine: A Technical Overview. In *Science, Technology and the Handicapped* (M. R. Redden and W. Schwandt, editors), pp. 3–11, Washington: American Association for the Advancement of Science, 1976. Report 76-R-11.

Samuel C. Lee, Shilin Xu, and Bailing Guo. Microcomputer-generated Chinese Speech. *Computer Processing of Chinese and Oriental Languages*, Vol. 1, No. 2, pp. 87–103, 1983.

Lin-Shan Lee, Chiu-Yu Tseng, and Ming Ouh-young. The Synthesis Rules in a Chinese Text-to-Speech System. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 37, No. 9, pp. 1309–1320, 1989.

Lin-Shan Lee, Chiu-Yu Tseng, and Ching-Jiang Hsieh. Improved Tone Concatenation Rules in a Formant-based Chinese Text-to-Speech System. *IEEE Transactions on Speech and Audio Processing*, Vol. 1, No. 3, pp. 287–294, 1993.

S. M. Lei. *Digital Synthesis of Mandarin Speech Using its Special Characteristics.* Master's thesis, National Taiwan University, 1982.

Mark Y. Liberman and Janet B. Pierrehumbert. Intonational Invariance under Changes in Pitch Range and Length. In *Language Sound Structure* (M. Aronoff and R. Oehrle, editors), pp. 157–233, Cambridge, Massachusetts: M.I.T. Press, 1984.

Philip Lieberman. *Intonation, Perception, and Language.* MIT Press, Cambridge, Massachusetts, 1967.

Wen C. Lin and Tzien-Tsai Luo. Synthesis of Mandarin by Means of Chinese Phonemes and Phonemes-pairs (JIFH). *Computer Processing of Chinese and Oriental Languages*, Vol. 2, No. 1, pp. 23–35, 1985.

Chi-Shi Liu, Ming-Fei Tsai, Yen-Huei Hsyu, Chan-Chou Lu, and Shiow-Min Yu. Yi Xianxing Yuce Bianma Wei Hechengqi de Zhongwen Wenju Fan Yuyin Xitong (A Chinese Text-to-Speech Based on LPC Synthesizer). *(TL Technical Journal)*, Vol. 19, No. 3, pp. 269–285, 1989.

Chi-Shi Liu, Gwo-Hwa Ju, Wern-Jun Wang, Hsiao-Chuan Wang, and Wen-Hsing Lai. A New Speech Synthesizer for Text-to-Speech System Using Multipulse Excitation with Pitch Predictor. In *International Conference on Computer Processing of Chinese and Oriental Languages*, pp. 205–209, 1991.

W. K. Lo and P. C. Ching. Phone-based Speech Synthesis with Neural Network and Articulatory Control. In *Proceedings of ICSLP 96*, 1996.

Shinan Lü, Shiqian Qi, and Jialu Zhang. Hecheng Yanyu Zirandu de Yanjiu (Study on the Naturalness of Synthetic Speech). *(Acta Acustica)*, Vol. 19, No. 1, pp. 59–65, 1994.

Yuhang Mao, Jinfa Huang, and Guozhen Zhang. A Chinese Mandarin Speech Output System. *European Conference on Speech Technology*, Vol. 2, pp. 87–92, 1987.

Mehryar Mohri and Richard Sproat. An Efficient Compiler for Weighted Rewrite Rules. In *34th Annual Meeting of the Association for Computational Linguistics*, (Morristown), pp. 231–238, Association for Computational Linguistics, 1996.

K. Nakata and T. Mitsuoka. Phonemic Transformation and Control Aspects of Synthesis of Connected Speech. *Journal of Radio Research Laboratories*, Vol. 12, pp. 171–186, 1965.

Ming Ouh-Young, Chin-Jiang Shie, Chiu-Yu Tseng, and Lin-Shan Lee. A Chinese Text-to-Speech System Based upon a Syllable Concatenation Model. In *IEEE ICASSP*, pp. 2439–2442, 1986.

Ming Ouh-Young. *A Chinese Text-to-Speech System.* Master's thesis, National Taiwan University, 1985.

Fernando Pereira and Michael Riley. Speech Recognition by Composition of Weighted Finite Automata. CMP-LG archive paper 9603001, 1996.

Fernando Pereira, Michael Riley, and Richard Sproat. Weighted Rational Transductions and their Application to Human Language Processing. In *ARPA Workshop on Human Language Technology*, pp. 249–254, Advanced Research Projects Agency, March 8–11 1994.

Gordon E. Peterson, William S-Y. Wang, and Eva Sivertsen. Segmentation Techniques in Speech Synthesis. *Journal of the Acoustical Society of America*, Vol. 30, pp. 739–742, 1958.

Janet Pierrehumbert and Mary Beckman. *Japanese Tone Structure*. The MIT Press, Cambridge, Massachussetts, 1988.

Janet Pierrehumbert and David Talkin. Lenition of /h/ and Glottal Stop. In *Papers in Laboratory Phonology II* (G. Docherty and R. Ladd, editors), pp. 90–116, Cambridge University Press, 1992.

Janet Pierrehumbert. *The Phonology and Phonetics of English Intonation*. PhD thesis, MIT, 1980.

Ralph K. Potter, George A. Kopp, and Harriet C. Green. *Visible Speech*. van Nostrand, New York, 1947.

S. Öhman. Word and Sentence Intonation, A Quantitative Model. Technical report, Department of Speech Communication, Royal Institute of Technology (KTH), 1967.

Delin Qin and N. C. Hu. A New Method of Synthetic Chinese Speech on a Unlimited Vocabulary. In *Cybernetics and Systems '88, Proceedings of the Ninth European Meeting on Cybernetics and Systems Research*, pp. 1223–1230, 1988.

Hongmo Ren. Linguistically Conditioned Duration Rules in a Timing Model for Chinese. In *UCLA Working Papers in Phonetics 62* (Ian Maddieson, editor), pp. 34–49, UCLA, 1985.

Jiong Shen. Beijinghua Shengdiao de Yinyu he Yudiao (Pitch Range and Intonation of the Tones of Beijing Mandarin). In *Beijing Yuyin Shiyan Lu (Acoustic Studies of Beijing Mandarin)*, pp. 73–130, Beijing University Press, 1985.

Xiao-Nan Susan Shen. *The Prosody of Mandarin Chinese*. University of California Press, 1990.

Xiao-Nan Susan Shen. Tonal Coarticulation in Mandarin. *Journal of Phonetics*, Vol. 18, pp. 281–295, 1990.

Bo Shi. A Chinese Speech Synthesis-by-rule System. In *Speech, Hearing and Language: Work in Progress, U.C.L. Number 3*, pp. 219–236, University College London, 1989.

C. T. Shie. *A Chinese Text-to-Speech System Based on Formant Synthesis*. Master's thesis, National Taiwan University, 1987.

Chilin Shih and Benjamin Ao. Duration Study for the AT&T Mandarin Text-to-Speech System. In *Proceedings of The Second ESCA/IEEE Workshop on Speech Synthesis*, (New Paltz, NY, USA), pp. 29–32, ESCA/AAAI/IEEE, 1994.

Chilin Shih and Benjamin Ao. Duration Study for the Bell Laboratories Mandarin Text-to-Speech System. In *Progress in Speech Synthesis* (Jan van Santen, Richard Sproat, Joseph Olive, and Julia Hirschberg, editors), New York: Springer, 1996.

Chilin Shih and Mark Y. Liberman. A Chinese Tone Synthesizer. Technical report, AT&T Bell Laboratories, 1987.

Chilin Shih. *The Prosodic Domain of Tone Sandhi in Chinese*. PhD thesis, University of California, San Diego, 1986.

Chilin Shih. The Phonetics of the Chinese Tonal System. Technical report, AT&T Bell Laboratories, 1987.

Chilin Shih. Tone and Intonation in Mandarin. In *Working Papers of the Cornell Phonetics Laboratory, Number 3: Stress, Tone and Intonation*, pp. 83–109, Cornell University, 1988.

Chilin Shih. Study of Vowel Variations for a Mandarin Speech Synthesizer. In *EUROSPEECH '95*, pp. 1807–1810, 1995.

Richard Sproat and Michael Riley. Compilation of Weighted Finite-State Transducers from Decision Trees. In *34th Annual Meeting of the Association for Computational Linguistics*, (Morristown), pp. 215–222, Association for Computational Linguistics, 1996.

Richard Sproat, Chilin Shih, William Gale, and Nancy Chang. A Stochastic Finite-State Word-Segmentation Algorithm for Chinese. *Computational Linguistics*, Vol. 22, No. 3, 1996.

Richard Sproat. A Finite-State Architecture for Tokenization and Grapheme-to-Phoneme Conversion for Multilingual Text Analysis. In *Proceedings of the EACL SIGDAT Workshop* (Susan Armstrong and Evelyne Tzoukermann, editors), (Dublin, Ireland), pp. 65–72, Association for Computational Linguistics, 1995.

Richard Sproat. Multilingual Text Analysis for Text-to-Speech Synthesis. In *Proceedings of the ECAI-96 Workshop on Extended Finite State Models of Language*, (Budapest, Hungary), European Conference on Artificial Intelligence, 1996.

J. Q. Stewart. An Electrical Analog of the Vocal Tract. *Nature*, Vol. 110, pp. 311–312, 1922.

Ching Y. Suen. Computer Synthesis of Mandarin. In *IEEE ICASSP*, pp. 698–700, 1976.

Chiu-Yu Tseng. *An Acoustic Phonetic Study on Tones in Mandarin Chinese*. PhD thesis, Brown University, 1981.

Jan P. H. van Santen. Assignment of Segmental Duration in Text-to-Speech Synthesis. *Computer Speech and Language*, Vol. 8, No. 2, pp. 95–128, 1994.

Michelle Q. Wang and Julia Hirschberg. Automatic Classification of Intonational Phrase Boundaries. *Computer Speech and Language*, Vol. 6, pp. 175–196, 1992.

Hsiao-Chuan Wang and Tai-Hwei Hwang. An Initial Study on the Speech Synthesis of Taiwanese Hokkian. *Computer Processing of Chinese and Oriental Languages*, Vol. 7, No. 1, pp. 21–36, 1993.

Wern-Jun Wang, Wen-Hsing Lai, Gwo-Hwa Ju, Chun-Hsiao Lee, and Chi-Shi Liu. Knowledge-based Prosodic Rule Generation in Mandarin Synthesizer. In *Workshop Notes of IEEE International Workshop on Intelligent Signal Processing and Communication Systems*, pp. 176–181, 1992.

Reiner Wilhelms-Tricarico and Joseph Perkell. Biomechanical and Physiologically Based Speech Modeling. In *Progress in Speech Synthesis* (Jan van Santen, Richard Sproat, Joseph Olive, and Julia Hirschberg, editors), pp. 221–234, New York: Springer, 1996.

C. H. Wu, C. H. Chen, and S. C. Juang. Yi CELP Wei Jichu zhi Wenju Fan Yuyin Zhong Yunlu Xunxi zhi Chansheng Yu Tiaozheng (An CELP-based Prosodic Information Modification and Generation of Mandarin Text-to-Speech). In *ROCLING VIII*, pp. 233–251, 1995.

Jun Xu and Baozong Yuan. New Generation of Chinese Text-to-Speech System. In *Proceedings TENCON '93, IEEE Region 10 Conference on Computer, Communication, Control, and Power Engineering*, pp. 1078–1081, 1993.

Yi Xu. *Contextual Tonal Variation in Mandarin Chinese*. PhD thesis, The University of Connecticut, 1993.

Shun-an Yang and Yi Xu. An Acoustic-phonetic Oriented System for Synthesizing Chinese. *Speech Communication*, Vol. 7, pp. 317–325, 1988.

David Yarowsky. Homograph Disambiguation in Text-to-Speech Synthesis. In *Progress in Speech Synthesis* (Jan van Santen, Richard Sproat, Joseph Olive, and Julia Hirschberg, editors), pp. 159–174, New York: Springer, 1996.

Jialu Zhang, Shiqian Qi, and Ge Yu. Assessment Methods of Speech Synthesis Systems for Chinese. In *Proceedings of the XIII International Congress of Phonetic Sciences*, pp. 206–209, 1995.

Jialu Zhang. Acoustic Parameters and Phonological Rules of a Text-to-Speech System for Chinese. In *IEEE ICASSP*, pp. 2023–2026, 1986.

Jialu Zhang. Hanyu Yuyin Xingneng Pingce Jieguo (Results of Chinese Speech Synthesizer Evalution). *(China Computerworld)*, pp. 175, 179, March 25 1996.

Qiaofeng Zhou and Lianhong Cai. Simulation of Stress in Chinese TTS System. Unpublished manuscript, 1995.

Ke-Cheng Zhou and Trevor Cole. A Chip Designed for Chinese Text-to-Speech Synthesis. *Journal of Electrical and Electronics Engineering, Australia*, Vol. 4, No. 4, pp. 314–318, 1984.

Ke-Cheng Zhou. A Chinese Text-to-Speech Synthesis System Using the Logarithmic Magnitude Filter. *Journal of Electrical and Electronics Engineering, Australia*, Vol. 6, No. 4, pp. 270–274, 1986.