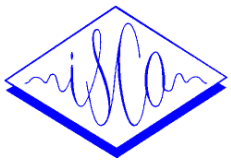

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

STUDY OF VOWEL VARIATIONS FOR A MANDARIN SPEECH SYNTHESIZER

Chilin Shih

AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, New Jersey 07974 USA

ABSTRACT

This paper reports the results of an acoustic study of Mandarin Chinese that was carried out for the AT&T Mandarin text-to-speech system. We present the optimal classification of vowels for the purpose of the synthesizer, and discuss some coarticulation effects and their implications for the collection of acoustic inventory elements. We are able to achieve excellent speech quality with a diphone-based concatenative system.

1. INTRODUCTION

Smooth connection and compatible direction of formant trajectories are two crucial issues that need to be addressed to insure the success of a concatenative speech synthesis. It is apparent why smooth connection is necessary. In a hypothetical case depicted in Figure 1, the F2 of the diphone unit *m-a* is not compatible with that of the diphone unit *a-t*. Connection like this causes unpleasant glitches in the synthesizer. A perfect match in the formant space, however, is not sufficient, Figure 2 depict another hypothetical case where the formant values of the two units match, but not the direction of the formant trajectories. Connection like this may lead to the perception of two sounds when one is intended. It is therefore important to have a good understanding of the acoustic properties of the speech sounds and the coarticulation effects of a given languages to avoid the two problems described above.

In the following we describe the results of such a study on Mandarin Chinese. Averaged formant values of vowels glides, and diphthongs are given in Sections 3, 4 and 5 respectively. Coarticulation effects will be discussed in Section 6.

2. DATA

Nonsense words covering all Mandarin vowels in all possible diphone contexts were embedded in frame sentences and were recorded by one male Beijing Mandarin speaker. There were three types of frame sentences to allow key words to be placed in the initial, medial, and final positions of a sentence. More than 3000 tokens were collected and analyzed. Values of the first three formants are taken at 20%, 35%, 50%, 65% and 80% of the vowel duration. The first and last 20% of the vowel duration are excluded from the study because we in general avoid those region while selecting cutting point of the acoustic inventory elements. The database is rich in contextual effects and allows us to systematically study both the magnitude and the duration of coarticulation effects on vowels.

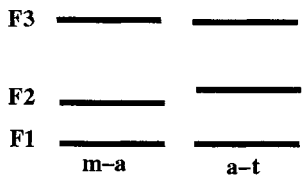

Figure 1: Incompatible formant values

3. MONOPHTHONGS

The vowel classification of this study is similar to the phonetic description of [1, 6], and the formant values conform in general with previous acoustic studies of Mandarin vowels [3, 8]. Figure 3 shows the median values of the first three formants (represented as 1, 2, and 3 respectively) of 11 monophthongs. There are three high vowels [i], [U] and [u], representing the front unrounded, the front rounded, and the back rounded vowels respectively. There are also two apical vowels [%] and [$], the former only co-occur with retroflex consonants and the later with dental consonants; the tongue positions inherit those of the preceding consonants. [R] is a heavy retroflex vowel. As expected, it has the characteristic lowering of F3 of a retroflex sound. Moreover, this is a very stable vowel due to a phonotactic constraint that it cannot co-occur with any other sound in the same syllable, which, if present, is the primary cause of coarticulation effects. The mid vowels include [e], [E], and [o], representing the mid front vowel, the schwa, and the mid back vowel respectively. [@] is a variant of [a] that

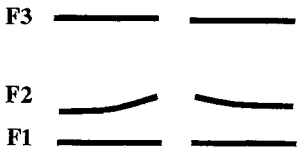

Figure 2: Incompatible formant trajectories

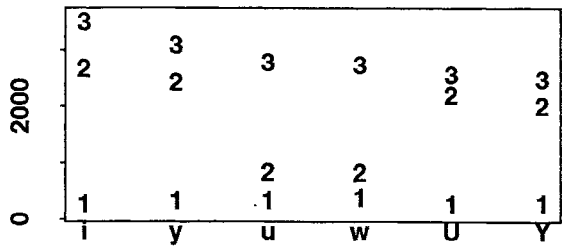

Figure 4: Formant values of glides and high vowels

is fronted and raised in the environment of a following alveolar nasal. This sound is treated as distinct from [a] in the synthesizer because of the big difference in F2 between the two sounds. Distinguishing [%] from [$] and [@] from [a] allow us to avoid the problem of incompatible formants depicted in Figure 1.

4. GLIDES

One issue about glide that is relevant to speech synthesis is whether they are sufficiently similar to high vowels that one can used high vowels in their places. Doing so has the advantage of reducing the number of necessary acoustic inventory elements (in our cases, they are primarily diphones). There are three on-glides in Mandarin, [y], [w], [Y], representing the front unrounded, the back rounded, and the front rounded glide respectly. The formant values of the glides are shown together with corresponding high vowels in Figure 4. Again, the median values of the first three formants are plotted as 1, 2, and 3 respectively. the front unrounded [y] has a very different formant structure than its corresponding high vowel [i], and the front rounded pair [Y] and [U] are somewhat different. In both cases, the on-glides have lower F2 and F3. The back rounded glide [w], on the other hand, has very similar formant structure as the back vowel [u]. Currently we do not merge glides with vowels in the synthesizer, but apparently one can at least merge [w] with [u] without much ill-effect.

5. DIPHTHONGS

Along the same line as the issue of the glide, a question of interest to the application of speech synthesis is whether diphthongs can be treated as combinations of vowels nucleus and off-glides. If this is possible one could save quite a number of acoustic inventory elements without sacrificing speech quality. Unfortunately, our study shows that, with the exception of [O], the vowel nucleus of a diphthong is typically different from the corresponding monophthong, which leads to our decision to represent diphthongs as whole units, rather than as combinations of vowels and off-glides. Figures 5, 6, 7, and 8 show the formant trajectories of diphthongs [I] *(ay)*, [W] *(aw)*, [A] *(ay)*, and [O] *(ow)* together with corresponding monophthongs.

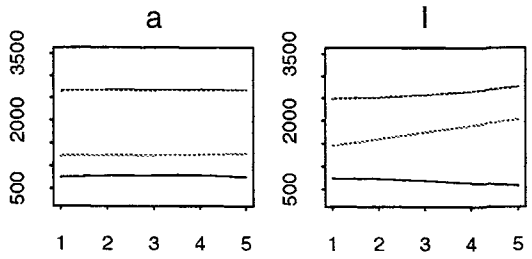

Figure 5: Formant trajectories of [a] and [I]

In Figure 5 the formant trajectories of diphthong [I] are compared to those of [a]. Note that the F2 of [I] is higher than that of [a] throughout the entire duration of the vowel. One way to compute the degree of formant discrepancies is to take root mean square (rms) difference of the three formants at selected points of the two sounds in question. The rms values comparing the 20% point of [I] with 20%, 35%, 50%, 65% and 80% points of [a] are 163, 169, 169, 167, 153 respectively. In contrast, the baseline rms values obtained by comparing a sound to itself are small. For example, the rms values comparing the 20% point with the other four positions of [a] are 21, 31, 32, and 29 respectively. What these numbers suggest is that the nucleus of [I], supposedly an [a] like vowel, is not similar to any cross-section of the monophthong [a]. The discrepancies between [W] and [a] as well as that between [A] and [e] are smaller, but still, even the smallest rms value, 77 between [W] and [a] and 47 between [A] and [e], are considerably bigger than the baseline value. Interestingly, in all three pairs so far the smallest rms values are obtained from the 20% point of the vowels. In other words, diphthongs [I], [W], and [A] are most similar to the beginning portion of the corresponding monophthongs.

Diphthong [O] presents a different picture. It is quite similar to the monophthong [o], the set of rms values comparing the 20% point of [O] with all five points of [o] is 111, 97, 79, 52, and 9. What these number show is that the formant values of [o] are changing and they become progressively similar to [O]. By the 80% point of [o] the formant values are identical to the 20% point of [O] in the sense that the rms value there is only a fraction of the baseline value obtained from [a].

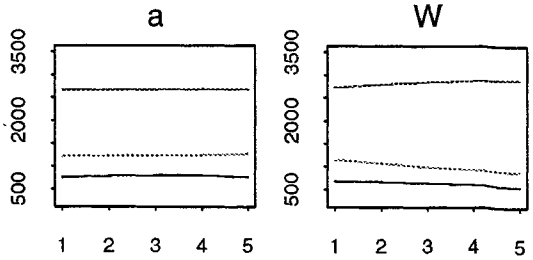

Figure 6: Formant trajectories of [a] and [W]

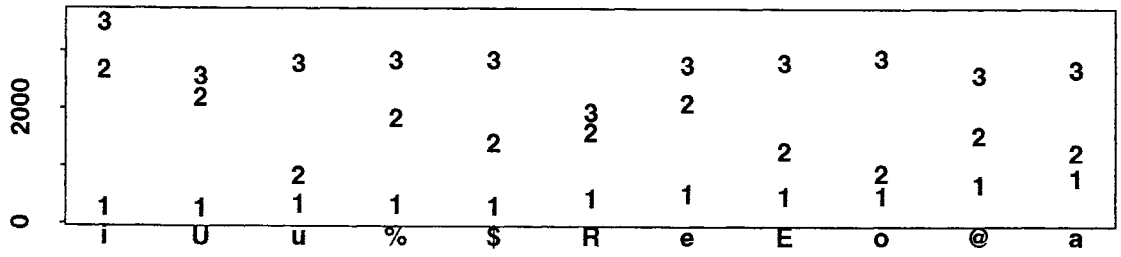

Figure 3: Formant values of Mandarin vowels

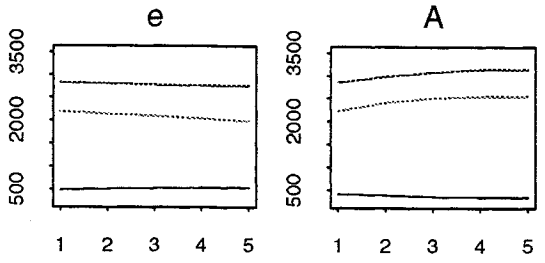

Figure 7: Formant trajectories of [e] and [A]

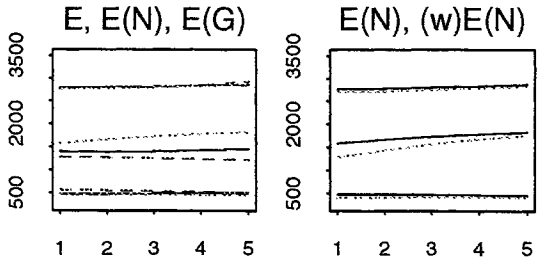

Figure 10: Coarticulation effects on [E]

6. COARTICULATION

Our database reveals many interesting coarticulation effects in Mandarin. Neighbouring sounds are the cause of coarticulation effects, however, different sounds may respond to the influence of surrounding sounds in different ways [2, 7, 4, 5].

When coarticulation effects persist through most of the duration of the vowel, or when there are significant changes in the direction of formant trajectories, it will be necessary to collect context-sensitive units to avoid connecting incompatible diphone units in the synthesizer. In the following we discuss two different patterns of coarticulation effects that result in different solutions for the speech synthesizer, the coarticulation effects of the schwa [E], and of the low vowel [a]. In the case of [E] it is necessary to split the phoneme by context to ensure smooth connection in the synthesizer. In the case of [a] no splitting is needed even though the coarticulation effects are quite strong.

Figure 10 compares the formant trajectories of [E] in different contexts. 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.

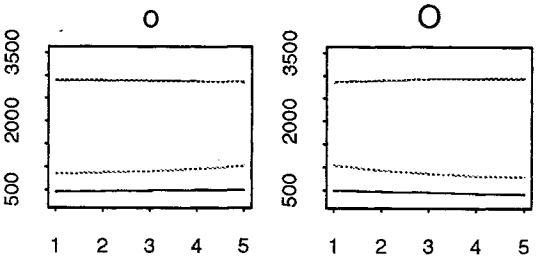

Figure 8: Formant trajectories of [o] and [O]

In the left-hand panel, solid lines are the average formant values of all the [E]'s in our database, dotted line are the average formant values of all the [E]'s in the [EN] context, where [N] is an alveolar nasal coda, and dashed lines are the average formant values of all the [E]'s in the [EG] context, where [G] is a velar nasal coda. The F2 of [E] in the [EN] context is higher than that of the average [E], while the F2 of [E] in the [EG] context is lower. The directions of the coarticulation effects are expected and are perfectly consistent with the anticipated tongue positions of the following sounds. However, the effects are already present in the first 20% of the vowel. In addition, the magnitude of the effects are strong. The difference in the F2 of [EN] and [EG] ranges from 300 Hz at the 20% point to 600 Hz at the 80% point.

The right-hand panel of Figure 10 show additional coarticulation effect caused by the back rounded onglide [w]. The solid lines plot the formant trajectories of all the [E]'s in the [EN] context, in contrast to the formant trajectories of [E]'s in the [wEN] context, which are plotted in dotted lines. A preceding [w] lowers the F2 of [EN], making it similar to the F2 of [E] in general. However, [wEN] is a poor source of general purpose diphone units, which typically have relatively flat F2 trajectories. Connecting a unit taken from [wEN] with rising F2 with other units with flat F2 will create a problematic situation described in Figure 2. Based on the coarticulation

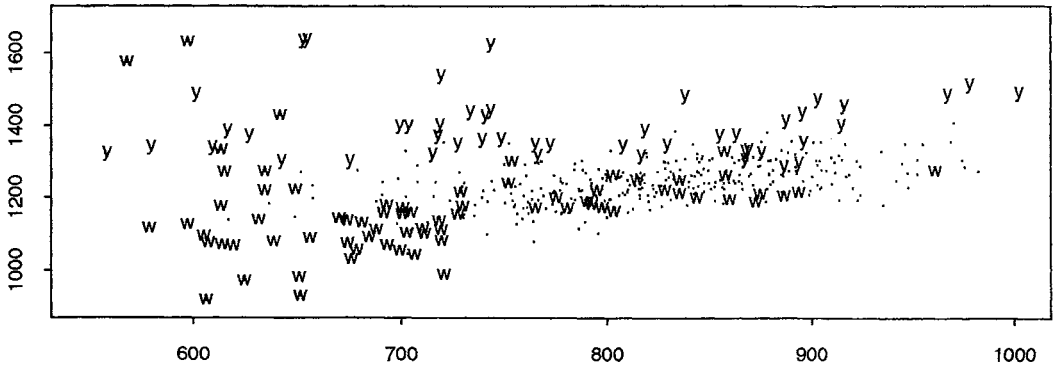

Figure 9: F1 and F2 of (y)a vs. (w)a

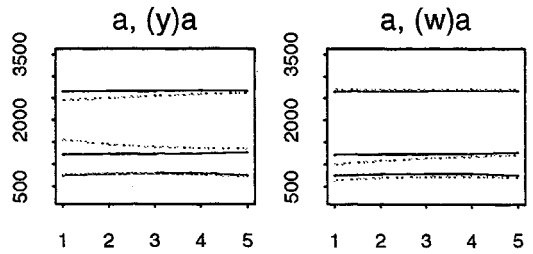

Figure 11: Coarticulation effects on [a]

effects observed above, we collected a triphone unit [wEN], and two sets of diphone units, one set taken from and used in the [N] context, the other set taken from and used in the rest of the contexts.

It should be noted that Mandarin [E] may received stress just as any other vowels do. [E] is not in itself a reduced vowel, although vowel reduction is realized as centralization and in extreme cases the resulting vowel quality is indistinguishable from [E]. The coarticulation effect discussed here is not a result of vowel reduction.

Not all strong coarticulation effects call for context-sensitive units. Plots in Figure 11 show the influence of [y] and [w] on the low vowel [a], respectively. The average formant values of [a] are plotted in solid lines, contrasting the dotted lines representing [a]'s in [ya] in the left panel and [wa] in the right panel. A preceding [y] has the effect of raising F2 and lowering F3, while a preceding [w] has the effect of lowering F1 and F2. The coarticulation effects of [y] and [w] are long-lasting, still present 65% into the vowel's duration, but after that the formant values come close to the averaged [a] values.

We examine the situation more closely in Figure 9, which is a scatter plot of F1 and F2 values taken from the mid point of all the [a]'s. The plotting symbol [y] represents the [a]'s preceded by [y], and the plotting symbol [w] represents the [a]'s preceded by [w]. All other preceding contexts are plotted with a period. The [y] and [w] populations are separated on the F2 dimension but some extreme values come close. The wide range on the F1 dimension is caused by other coarticulation effect, i.e., nearly all samples with F1 lower than 700 Hz are followed by [G]. The gap between the [y] and [w] population is also bridged by the rest of the [a] which lies in between them. The merging formant values and the leveling of the formant trajectories suggest that, despite the strong coarticulation effect during the first half of the vowel, no context-sensitive diphones of [a] is needed for the observed differences.

7. CONCLUSIONS

The current study provides necessary information that enable us to devise a proper classification of Mandarin vowels for a concatenative speech synthesizer. Contexts that cause strong coarticulation effects are identified and studied, which allow us to make informed choices on the selection of acoustic inventory elements for the Mandarin synthesizer.

8. REFERENCES

1. Y. R. Chao. *A grammar of spoken Chinese.* University of California Press. Berkeley 1968.
2. C. A. Fowler. *Coarticulation and theories of extrinsic timing.* Journal of Phonetics 8. 1980, pp.113-133.
3. J. M. Howie. *The vowel and tones of Mandarin Chinese: Acoustical measurements and experiments.* Doctoral dissertation. Indiana University 1970.
4. R. D. Kent and F. D. Minifie. *Coarticulation in recent speech production models.* Journal of Phonetics 5. 1977, pp.115-117.
5. B. Lindblom. *Economy of speech gestures.* in P. F. MacNeilage ed. *The production of speech.* Springer-Verlag. New York 1983, pp.217-245.
6. J. Norman. *Chinese.* Cambridge University Press. Cambridge 1988.
7. D. H. Whalen. *Coarticulation is largely planned.* Journal of Phonetics 18. 1992, pp.3-35.
8. Z. J. Wu and M. C. Lin eds.: *Shiyan Yuyinxue Gaiyao [Experimental Acoustics].* Higher Education Press. Beijing 1989.