MODELING VOWEL DURATION FOR JAPANESE TEXT-TO-SPEECH SYNTHESIS

Jennifer J. Venditti [*][†] *and Jan P. H. van Santen* [*]

[*]Bell Labs – Lucent Technologies   [†]Ohio State University

ABSTRACT

Accurate estimation of segmental durations is crucial for natural-sounding text-to-speech (TTS) synthesis. This paper presents a model of vowel duration used in the Bell Labs Japanese TTS system. We describe the constraints on vowel devoicing, and effects of factors such as phone identity, surrounding phone identities, accentuation, syllabic structure, and phrasal position on the duration of both long and short vowels. A Sum-of-Products approach is used to model key interactions observed in the data, and to predict values of factor combinations not found in the speech database. We report root mean squared deviations between observed and predicted durations ranging from 8 to 15 ms, and an overall correlation of 0.89.

1. INTRODUCTION

Segmental durations in natural speech are highly context dependent. For example, the first /a/ in the Japanese word *kakaru* 'to hang up' is shorter than the second, which is in turn shorter than the /a/ in *suika* 'watermelon' when said in isolation. Phone identity, surrounding phone identities, and phrasal position are just a few of the factors which can exhibit influences on segment duration in natural speech. One of the goals of TTS systems is to provide an accurate estimate and model of this duration variation, based on factors that can be identified and computed from text. This paper presents an analysis of vowel durations in Tokyo Japanese read speech for implementation in the Bell Labs Japanese TTS system.

One of the main problems in investigating contextual effects on segmental durations is *data sparsity*. It is certainly impossible to measure durations of every factor combination (or "cell") in a given language, and due to the large number of very rare combinations, it is not sufficient to simply model only the most common combinations. However, in most cases these contextual effects are well-behaved in terms of *directional invariance*. That is, the direction of an effect of a given factor is the same in all cases (e.g. phrase-final vowels are longer, regardless of the vowel identity), and there are no reversals of such effects. Phenomena having this property can be described reasonably accurately with Sum-of-Products (SoP) models [10].

Sum-of-Products models are a generalization of additive and multiplicative models, and consist of sums of terms, each term itself being a multiplicative model for a subset of one or more factors. These models can capture any interaction pattern, provided it is well-behaved. The key advantage of SoP models over, for example, Classification and Regression Trees (e.g. [5]), is that they provide good estimates of missing combinations, via *interpolation*;

this interpolation rests on the fact that directional invariance implies that factor combinations are intrinsically ordered, so that values for missing combinations are constrained by observed values of adjacent combinations. We will show that the SoP model approach yields robust prediction of Japanese long and short vowel durations.

2. SPEECH DATABASE

For sentence selection, we used a set of approximately 34,000 sentences contained within a larger newspaper text database. The sentences were all 20-40 characters in length, and contained at least one comma. The phonetic makeup and prosodic groupings of each sentence were determined using text analysis methods developed at Bell Labs [7]. Each of the phones was coded by a vector of features (factors) describing the phone and the context in which it occurred. The factors (and levels of each factor) used in this study are listed below, and are among those known to affect segment durations in Japanese and other languages [1, 3, 6, 8, 9].

- **current phone identity:** This study examines short vowels (/a/, /i/, /u/, /e/, /o/) and long vowels (/A/, /I/, /U/, /E/, /O/).
- **preceding phone identity:** Voiceless stop, voiceless fricative/affricate, voiced stop, voiced fricative, flap, nasal, glide, vowel.
- **following phone identity:** Voiceless stop/affricate, voiceless fricative, voiced stop, voiced fricative, flap, nasal, glide, vowel.
- **left prosodic context:** The syllable is: major phrase (MaP)-initial, minor phrase (MiP)-initial, intonation phrase (IP)-initial, accentual phrase (AP)-initial, non-initial.
- **right prosodic context:** The syllable is: major phrase-final, minor phrase-final, intonation-phrase final, accentual phrase-final, non-final.
- **accent status:** The syllable is: accented (acc), downstep accented (dnstp), preceding an accent in an accented AP (pre), following an accent in an accented AP (post), in an unaccented AP (unacc).
- **syllable structure:** The syllable is: open (V, CV, CyV) or closed by a geminate or moraic consonant (VC, CVC, CyVC).
- **special morpheme status:** The phone is part of one of the following: copula /desu/, verbal ending /masu/, perfect marker /ta/, copula /da/, topic marker /wa/, particle /ga/, particle /to/, other case particles, or none of the above.

It is important to note that the prosodic categories *major phrase* and *minor phrase* differ from the standard usage in Japanese intonational phonology (e.g. [4]). We follow the conventions in our

text analysis module, which define a major phrase as a sequence of phones delimited by an orthographic period, and a minor phrase as that delimited by a comma. Each sentence in our database is comprised of exactly one major phrase, and two or more minor phrases. Intonational (aka. intermediate) phrasing and accentual phrasing were hand-coded from the spoken utterances, using the J_ToBI prosodic transcription system [11].

Factors were grouped together by those which were expected to interact (e.g. phone identity with previous context, etc.), producing a list of several sub-feature-vectors characterizing each phone. With the entire database coded as such, a greedy algorithm [10] was used to select the smallest set of sentences which completely cover the entire set of unique sub-feature-vectors occurring in the 34,000 sentence database. The selection process resulted in a total of 197 sentences covering this space.

The selected sentences were recorded by a male native speaker of Tokyo Japanese in a sound-attenuated room at Bell Labs, and were segmented by a trained human labeler. Coded factors of each phone were adjusted to reflect the actual production of the utterances.

3. ANALYSIS

The duration analysis and modeling were conducted using a statistical analysis package developed at Bell Labs [10]. In this paper, we report only on the analysis of vowel durations (6092 vowels total). A full description of our Japanese duration model, including consonants, can be found in [12].

Campbell [1] proposes that the sentence-final shortening and other durational effects described by Kaiki et al. [3] are due to an imbalance in their database involving special morphemes. To investigate this hypothesis further, we partitioned the vowel data into two subsets: vowels contained in one of the special morphemes listed above (1039 vowels), and all others (5053 vowels). Due to space constraints, the current paper will focus on the effects on vowels *not* contained within special morphemes. See [12] for an analysis of special morphemes.

3.1. Vowel devoicing

One very well-known effect on vowel duration in Japanese is the phenomenon of devoicing, which causes high vowels to be partially or totally devoiced when flanked by voiceless consonants (hereafter, the 'devoicing context'). In our database, totally-devoiced (hereafter 'devoiced') vowels are marked by having 0 ms duration. Figure 1 shows the percentage of devoiced vowels in four voicing environments.

While a majority of the devoiced vowels are high vowels surrounded by voiceless consonants, this is neither a necessary nor sufficient criterion. Over 20% of high vowels in the unvoi_voi environment are devoiced. The fact that vowels are not devoiced in the voi_unvoi environment indicates that the voicelessness of the preceding consonant has more influence on devoicing than that of the following consonant. In addition, 20% of high vowels in the canonical devoicing environment are not devoiced, suggesting that there are other factors at work in preventing total-devoicing.
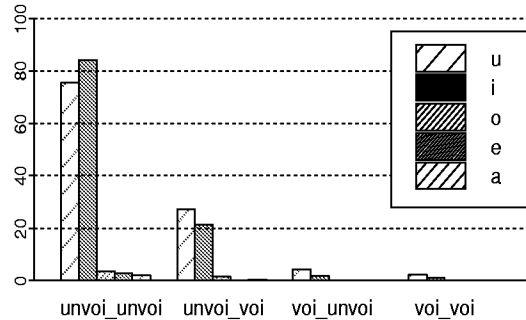

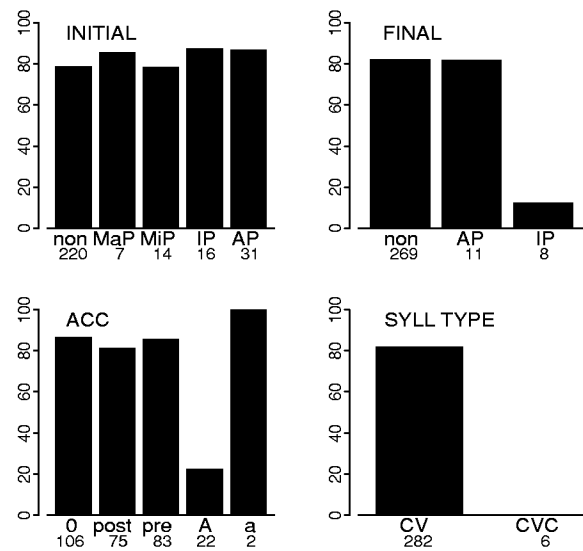**Figure 1:** Percentage of devoiced vowels by voicing environment and vowel identity.

Figure 2: Percentage of devoiced vowels in canonical devoicing context. Total number of observations is given below each type.

Figure 2 shows the percentages of devoiced vowels in the devoicing environment, arranged by factors coded in our database. There is a clear categorical distinction between vowels in this environment which tend to be devoiced, and those that do not. While the number of observations are admittedly low, we observe that the majority of IP-final, accented (see also [2]), and vowels followed by voiceless geminates tend not to be devoiced.

Based on these observations, we model the categorical nature of vowel devoicing in Japanese by setting durations of high vowels in the devoicing context to 0 ms, with the exception of the three categories with very low devoicing percentages shown in Figure 2. All remaining vowels are modeled using Sum-of-Products models, as described below.

3.2. Effects on short vowels

One of the contextual effects on Japanese vowel duration reported in the literature is that of prosodic position. Kaiki et al. [3] found accentual phrase-initial shortening, and AP-final and breath group-final (our MiP-final) lengthening effects. In addition, they
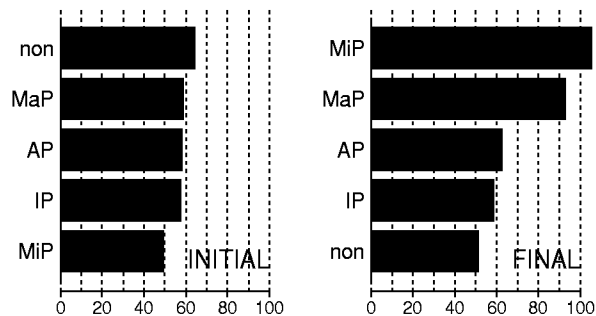
Figure 3: Corrected means (ms) of short vowels, arranged by prosodic position.

observed sentence-final shortening. Campbell [1] suggests that an imbalance in their speech corpus gives the false impression of shortening, due to the frequent occurrence of the short /-ta/ perfect marker in this position, and the long /wa/ topic marker in MiP-final position. In the current analysis, vowels in special morphemes have been removed. Figure 3 shows the *corrected means* (means of the levels on the factor of interest which have been corrected for the effects of the remaining factors, using a multiple regression method [10]) for short vowels (4481 vowels) in the prosodic positions that were coded in our study.

We observe initial shortening in MiP-initial position (post-pausal but not utterance-initial), and lengthening in major/minor phrase-final position (pre-pausal). This lengthening effect runs contrary to Kaiki et al.'s sentence-final shortening. However, it is the case that, of the two phrase levels, MaP-final vowels are lengthened to a lesser degree. This effect cannot be accounted for by the confounds Campbell proposes, since special morpheme vowels are not included in this analysis. We conjecture that MiP-final vowels are lengthened to a greater extent in order to cue continuation, or some forward-looking function in the discourse.

Because of the confounding of initiality/finality and surrounding phones (i.e. silences), the short vowel data were divided into 4 subgroups for subsequent analysis: minor/major phrase-final (pre-pausal), minor/major phrase-initial (divided into #_ and #C_ cases), and non-initial/non-final. Figure 4 shows the corrected mean values for each of the other coded factors, for non-initial/non-final short vowels (3542 vowels).

In these plots, the distinctions in preceding and following phone identity have been reduced to voicing status. We find that vowels are shorter when preceded by voiceless consonants than by voiced consonants or flaps. Similar effects hold for the voicing of the following consonant, albeit to a lesser degree. This suggests that the preceding consonant has a greater effect on the vowel duration, which is consistent with the effects on devoicing shown in Figure 1. Non-downstepped accented vowels in our database are longer than those in other accent positions, with post-accentual vowels being shortest. As for syllable type, vowels are shortest when preceded by a /y/ consonant off-glide, and are longest in syllables closed by a geminate or moraic consonant. All else being equal, vowels in syllables containing onset consonants are shorter than in their onset-less counterparts. These plots
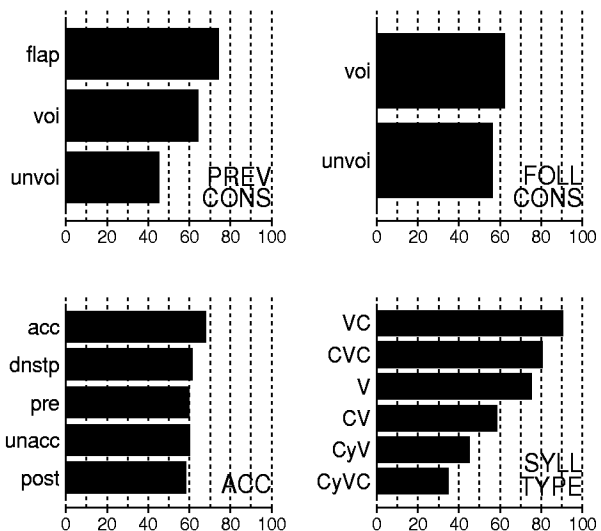

Figure 4: Corrected means (ms) of short vowels, arranged by phonetic environment, accent status, and syllable type.
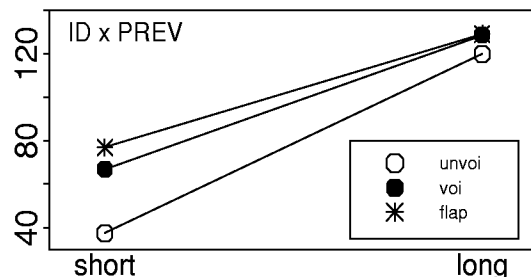

Figure 5: Two-way corrected means (ms) showing the interaction of vowel length with preceding consonant identity.

of non-initial/non-final vowels are representative of the effects in the other vowel subgroups as well.

3.3. Effects on long vowels

Contextual effects on long vowel durations in our corpus are comparable to the results for short vowels presented above. However, the magnitude of the effects are not as great. Figure 5 shows an example of one such interaction. Both short and long vowels are shorter when preceded by a voiceless consonant than by a voiced consonant or flap, though the effect is greater for short vowels.

4. MODEL

Given that the contextual effects on long and short vowels are comparable but of different magnitudes, we collapse both long and short vowels and apply a Sum-of-Products approach, which is able to incorporate the observed interactions. A separate SoP model was applied to each of the four vowel subgroups, with the exact nature of the model depending on a careful analysis of which factors interact in each subgroup. For example, to estimate

vowel category	N (cells)	RMS dev.	corr.
non-init/non-final	740	9 ms	.87
final	78	8 ms	.94
initial (#C_)	176	13 ms	.90
initial (#_)	35	15 ms	.85
all vowels	1029	—	.89

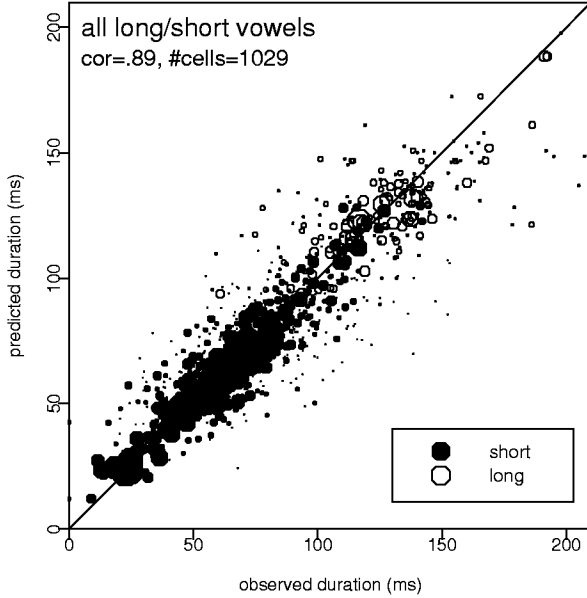**Table 1:** Results of SoP model parameter estimation for both long and short vowels combined.

Figure 6: Observed vs. predicted values for all vowels.

durations in the non-init/non-final subgroup, we use the following model, where $S_{i,j}$ are parameters corresponding to specific factor levels (*j*) in specific terms (*i*).

$$DUR(id, prev, foll, left\_pos, right\_pos, acc, syll) =$$
$$S_{1,1}(id) + [S_{2,1}(id) \times S_{2,2}(prev)] + [S_{3,1}(id) \times S_{3,3}(foll)] +$$
$$S_{4,4}(left\_pos) + [S_{5,1}(id) \times S_{5,5}(right\_pos)] +$$
$$[S_{6,1}(id) \times S_{6,6}(acc)] + [S_{7,3}(foll) \times S_{7,7}(syll)]$$

Table 1 gives the results of the parameter estimations. The number of observed factor combinations (N), the root mean squared (RMS) deviation between observed and predicted values, and the correlation coefficient are given for each vowel subgroup.

Figure 6 shows a scatter plot of observed and predicted values for all of the vowels analyzed in this study. The size of the plot characters are roughly proportional to the number of observations in each cell. Cells with many observations lie close to the $x=y$ diagonal line, while points lying away from the diagonal tend to be cells containing only one observation. The overall correlation is 0.89.

5. SUMMARY

This paper presents a quantitative model of Japanese vowel durations, for use in text-to-speech synthesis. We discuss constraints on vowel devoicing, and effects of phone identity and contextual factors on durations of long and short vowels. Sum-of-Products models are used to model the key interactions observed in the data, and to predict missing values based on interpolation. The result is a robust prediction of segmental duration, while at the same time being phonetically-motivated in that qualitative analysis of the data drives the modeling process.

6. ACKNOWLEDGMENTS

7. REFERENCES

1. Campbell, N. Segmental elasticity and timing in Japanese speech. In *Speech Perception, Production, and Linguistic Structure*, Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka, Eds. IOS Press, 1992, pp. 403–418.

2. Han, M. *Japanese Phonology: An Analysis Based upon Sound Spectrograms*. PhD thesis, University of Texas at Austin, 1961.

3. Kaiki, N., Takeda, K., and Sagisaka, Y. Statistical analysis for segmental duration rules in Japanese speech synthesis. In *Proceedings of the 1990 International Conference on Spoken Language Processing* (Kobe, Japan, 1990), pp. 17–20.

4. Kubozono, H. *The Organization of Japanese Prosody*. Kuroshio Publishers, 1993.

5. Riley, M. D. Tree–based modelling of segmental durations. In *Talking Machines: Theories, Models, and Designs*, G. Bailly, C. Benoit, and T. R. Sawallis, Eds. Elsevier Science Publishers, 1992, pp. 265–273.

6. Sagisaka, Y. On the modeling of segmental duration control. In *Speech Perception, Production, and Linguistic Structure*, Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka, Eds. IOS Press, 1992, pp. 451–455.

7. Sproat, R., Möbius, B., Maeda, K., and Tzoukermann, E. Multilingual text analysis. In *Multilingual Text–to–Speech Synthesis*, R. Sproat, Ed. Kluwer Academic Publishers, 1998.

8. van Santen, J. P. H. Contextual effects of vowel duration. *Speech Communication 11* 1992, 513–546.

9. van Santen, J. P. H. Assignment of segmental duration in text–to–speech synthesis. *Computer Speech and Language 8* 1994, 95–128.

10. van Santen, J. P. H., and Sproat, R. Methods and tools. In *Multilingual Text–to–Speech Synthesis*, R. Sproat, Ed. Kluwer Academic Publishers, 1998.

11. Venditti, J. J. Japanese ToBI labelling guidelines, 1995. [http://ling.ohio-state.edu/Phonetics/J_ToBI/jtobi_homepage.html].

12. Venditti, J. J., and van Santen, J. P. H. Modeling segmental durations for Japanese text-to-speech synthesis. In *Proceedings of the 3rd ESCA Workshop on Speech Synthesis* (Jenolan Caves, Australia, 1998).