MULTI-LINGUAL DURATION MODELING

Jan van Santen, Chilin Shih, Bernd Möbius, Evelyne Tzoukermann, and *Michael Tanenblatt*

Lucent Technologies – Bell Labs, 600 Mountain Avenue, Murray Hill, NJ 07974, USA
http://www.bell-labs.com/project/tts/

ABSTRACT

Controlling timing in text-to-speech synthesis systems is complicated, because there are many contextual factors that affect timing; moreover, which factors matter and what their precise effects are varies among languages. We describe here a language-independent approach for duration control. At run time, a language-independent timing module accesses language-specific tables. These tables specify which sub-classes of the feature space (i.e., all combinations of context and phone identity) are homogeneous in the specific sense that the same factors have similar effects on the cases in a sub-class. Within a sub-class, durations are modeled by simple arithmetic models such as multiplicative, additive, or – more generally – sums-of-products models. Exploratory statistical methods (supervised) and parameter estimation techniques (unsupervised) are used for table construction.

1. INTRODUCTION

In natural speech, durations of a given phone in different contexts can vary over as much as an order of magnitude. Much of this variation is not random, and correlates systematically with context and phone identity. These large durational effects obviously must be mimicked for speech generated by a text-to-speech system to sound natural.

After a brief review of duration modeling where we discuss the problems inherent in accurate duration prediction, we focus on the special challenges faced when a duration component has to be maximally "language independent". By language independence we mean the following. First, at run time the same executable is used for all languages; it reads tables containing all the language-specific information. Second, the information stored in these tables is largely derived by automatic methods from a segmented, labeled natural speech data base. Their construction does not significantly increase the amount of language-dependent manual labor required for construction of the other components of a synthesis system for a given language, such as the pronunciation, intonation, and synthesis components.

2. MODELING OF TIMING: THEORY

Segmental Duration in Perspective.
Like almost all timing modules, our module controls timing through segmental duration. However, we expect that in the future controlling timing through segmental duration will give way to controlling timing through sub-segmental time-warping. In fact, given the limits on how many coarticulatory phenomena can be captured using even very large acoustic unit inventories, we expect that ultimately some or many of these phenomena have to be modeled explicitly by dissecting and re-constituting acoustic units at run time; the trajectories of the parameters produced by dissection will be manipulated with separate time warps.

Properties of Segmental Duration.
Accurate prediction of segmental duration from text is difficult for two reasons. First, the total feature space of context-phoneme combinations is very large [12]. As a result, data based approaches to the construction of duration components must be able to handle cases not seen in the training data base (*generalization*). It has been shown that general purpose prediction techniques such as classification and regression trees (*CART*) have poor generalization properties, whereas methods based on quasi-linear regression perform much better [4].

Second, some factors affecting duration *interact*: the *magnitude* of their effects are affected by other factors. This makes application of standard linear regression problematic. For example, the effect on vowel duration of postvocalic voicing is much larger (measured either in milliseconds or as a fraction) for utterance-final syllables than for utterance-medial syllables [3].

However, what makes duration modeling easier is that many factors have the property of *directional invariance*. This concept is illustrated, for example, by the fact that, holding all else constant, in English a stressed vowel is longer than an unstressed vowel. In other words, the direction of the effects of a factor (such as stress) is unaffected by the remaining factors (possibly restricted to some subset of the total feature space, such as vowels). This means that while factors interact in that the *magnitude* of their effects are affected by other factors, they do not interact in that the *direction* of their effects is affected. As far as we know, in English all factors affecting vowel duration (i.e., vowel identity, identity of the postvocalic consonant, identity of the prevocalic consonant, location in the word and phrase, word accent, and sentence accent) have the property of directional invariance [8].

As a result of these considerations, we use an approach that captures interactions with highly constrained quasi-

linear models, which we call sums-of-products models [10]. A sums-of-products model is an equation whose parameters correspond to factor levels (e.g., "unstressed" is a level on the stress factor); these parameters are combined by taking products of sub-groups of parameters, and then adding the products. For example:

$$\mathrm{DUR}(\mathrm{Vowel}:/e/, \mathrm{Next}:Voiced, \mathrm{Loc}:Final) = \alpha(/e/) + \delta(Final) + \beta(Voiced) \times \gamma(Final)$$

This equation states that the duration of the vowel $/e/$ followed by a voiced consonant in utterance-final position is given by taking the intrinsic duration of the vowel [$\alpha(/e/)$], adding a millisecond amount [$\delta(Final)$] for being utterance final, and finally adding the effect of post-vocalic voicing [$\beta(Voiced)$] modulated by utterance-finality [$\gamma(Final)$].

We make the claim that *sums-of-products models have the property that they can capture directionally invariant interactions using very few parameters.* General-purpose prediction techniques need more parameters, and cannot capture deep regularities in the data such as directional invariance. For example, it is not impossible that – with the right kind of "holes" in the training data – CART predicts certain vowels to be shortened instead of lengthened by stress [11]. We conjecture that this is the key reason underlying Maghbouleh's [4] results.

To apply sums-of-products models to duration, we construct a tree such that the terminal nodes split the feature space into *homogeneous sub-classes*. The cases in a homogeneous sub-class have the property that *the same factors have similar effects, and hence can be modeled by one sums-of-products model.* For example, the sub-class consisting of voiceless stops in syllable onsets has the property that all 6 cases (closure regions and bursts of /t/, /p/, and /k/) are made longer by syllabic stress, are unaffected by the place of articulation of the following vowel, are shortened when preceded by a tautosyllabic voiceless fricative, and are shortened when followed by /l/. The sub-class of voiceless stops in syllable codas behaves differently; for example, syllabic stress has little or no effect, and stops in codas cannot be followed by tautosyllabic /l/. Although the magnitude of the effects of a factor in a sub-class varies over the cases (e.g., the effect of stress is larger for /t/ than for /p/), the direction of the effects (making durations longer vs. shorter vs. no effect) is the same. Homogeneity of a subclass translates into the ability to model the cases in that subclass with one sums-of-products model (i.e., an equation such as above) and accompanying parameters [i.e., the actual numerical values of $\alpha(/e/)$, $\delta(Final)$, etc.].

Other Approaches.
Our approach contrasts with two standard approaches to duration modeling. One approach, already discussed, consists of using general purpose prediction methods. This approach is certainly language-independent as defined above. Because of our concerns about the lack of generalization ability of these methods, we do not use this approach.

The other approach, exemplified by MITalk [1], uses manually constructed duration rules. This approach is language independent because the rules are stored in tables in a language-independent format. However, their construction is manual, and there are good reasons to be concerned about the efficiency of the construction process, the accuracy of the modeling of interactions, and the validity of the parameter estimates [9, 11].

3. MULTI-LINGUAL DURATION MODELING

Construction of a duration module for a given language consists of the following steps:

Step 1: Data Base Construction. Elsewhere, we have discussed methods for optimal text selection [13]. These methods maximize the coverage of the feature space, by selecting a training text corpus from a large text corpus with "greedy" methods. This requires the availability of text analysis components for computing duration module input automatically.

It should be noted that due to the combinatorial complexity of any unrestricted language domain, even the most sophisticated text selection algorithms produce training text with disappointing coverage of the domain [12]. *Sparsity of the training corpus remains a central problem in duration data analysis.*

While speaker selection and recording is essentially unproblematic, segmentation still has to be done manually. Although progress has been made in automatic segmentation, automatically detected phone boundaries still tend to have large errors and – worse – systematic biases, and hence cannot be used for accurate timing studies, in particular studies of short segments such as schwa's, voiced stop bursts, flapped voiceless stops, and glides.

Step 2: Sub-Class Structure. Determination of the sub-class structure is usually straightforward, because the classification usually follows standard distinctions and is only two or three layers deep. In the American-English system, for example, the first branching is between vowels and consonants; the next is between intervocalic and non-intervocalic consonants. It is known from the literature that different factors play different roles in these classes, thereby preventing homogeneity. Thus, sub-class structure is based on common sense, the acoustic-phonetics literature, and on analysis of the homogeneity of the durations in sub-classes via model fitting.

Step 3: Factors and Models. The determination of which factors are relevant, which distinctions (factor levels) must be made on each factor, and which sums-of-products model to use, is done with exploratory data analysis methods. Standard statistical tests are used here, but

	Chinese	French	German
Number Cases	46,265	7,143	24,240
Number of Sub-classes	6	16	30
Number of Parameters	298	782	674
Correlation	0.872	0.847	0.896
Rms	0.026	0.025	0.019
Obs Mean	0.076	0.074	0.060
Obs Std Dev.	0.053	0.047	0.043

Table 1: Number of cases, terminal nodes, estimated parameters, correlation and root mean squared deviation of observed and predicted durations (Rms, in seconds), and mean and standard deviation of the observed durations (both in seconds).

we also use graphical means of displaying patterns in the data. In case the data are sparse, or very poorly balanced, standard statistical tests can produce artifacts; visual, phonetically informed inspection of data patterns is often the preferred mode in which these determinations are made.

A critical role is played by "piecewise multiplicative correction" [8]. Here, we change an N-factor data set into a two-factor data set by combining all but the factor of interest into a compound factor containing all combinations of levels on these remaining factors. We then apply the multiplicative model to these new factors, but by being able to drop the requirement that the parameters for the compound factor be estimable, parameter estimation is often possible. The estimated parameters correspond to the levels on the factor of interest, corrected for any effects of the remaining factors. These corrected marginal means are typically much more valid than the raw marginal means, which may be seriously biased by factor imbalances (see [8] for an example).

The decision on which factors to include in the analysis and which levels to distinguish on each factor is based on piecewise multiplicative correction, on understanding of the effects of sparsity in the data on parameter estimate reliability, and on acoustic-phonetics knowledge.

Model selection is likewise seriously limited by data sparsity. For the class of sums-of-products models, exhaustive methods exist for model selection [10], but these methods are rarely used because they require completely balanced data sets which, due to linguistic and practical constraints, can only be produced for small sets of factors. In practice, we fit a small number of models and evaluate their fit with standard statistical methods.

Step 4: Parameter Estimation. All model fitting is performed with least-squares estimation in the log duration domain. In cases with extreme sparsity, we use robust estimation.

Summary. Once a segmented database is available, our tools allow construction of a duration component in a few days. This is a small amount of time by comparison to construction of other components of a text-to-speech system (in particular the text analysis components), and also by comparison to older approaches where duration rules were crafted manually and tested perceptually (e.g., MITalk). Since construction of a text-to-speech system for a new language is in part a process of discovery, we consider the manual aspects of duration component construction of value in their own right, adding to the depth and breadth of the understanding of the language in question.

4. EXAMPLES OF MULTI-LINGUAL DURATION MODELING

The Bell Labs Text-to-Speech System has employed this procedure for all covered languages. Here we describe as examples Mandarin Chinese [6], French [7], and German [5].

Our system has a group of *core factors*, which is the union of all factors used in any of our languages. Naturally, this group expands when some highly language-specific factors such as Chinese tone are encountered. However, as the number of covered languages grows, we expect the group of core factors to become reasonably complete. For example, for several languages (Spanish, Romanian, German) no additional factors were required.

Among the core factors are the following. First, *phone identity factors*: Identity of the (1) current segment, (2) previous segment(s), and (3) next segment(s). Second, *stress-related factors*: (4) Degree of discourse prominence, and (5) Lexical stress. And third, *locational factors*: Location of the (6) segment in the syllable, (7) syllable in the word, (8) word in the phrase, (9) phrase in the utterance.

Mandarin Chinese. Training text was created by greedy text selection methods applied to a text corpus of 15,620 newspaper sentences.

We had to add three *tone-related factors* to the group of core factors: Identity of the current, previous, and following tone. A second Chinese-specific factor, Syllable type (*CG-vowel-C, CG-vowel, C-vowel-C, CG-diphthong, C-vowel, C-diphthong, vowel-C, vowel*, and *diphthong*; here, C is a non-glide consonant, and G a glide), was computable from the core factors.

A classification scheme was constructed with six subclasses: *vowels, fricatives, plosive bursts and aspirations, plosive closures, sonorants in syllable onsets*, and *sonorants in syllable codas*. This classification is exhaustive, because codas can only contain sonorants.

Among findings that we typically have not seen in other languages were the following. First, absence of utterance-final lengthening. Second, a compensatory effect where coda consonants where shortened by intrinsically long nuclei. Some of the overall statistics are shown in Table 1.

French. Similar to Chinese, the training text was created by greedy text selection methods applied to a text corpus of 15,000 newspaper sentences.

A factor specific to French is whether a consonant is created by *liaison*. Here, "les autres" (the others) must be pronounced /le **z**otr(ə)/ and not /le otr(ə)/, with the addition of an intervocalic consonant /z/. If the second word does not start with a vowel, the /z/ is not pronounced.

Issues of special interest are complete *schwa deletion* in words such as /guvɛrnmɑ̃/ ("gouvernement", or government), and the high degree of confounding of lexical stress (always on word final syllable unless its nucleus is a schwa) with intra-word location.

It was found that consonants created by liaison are shortened relative to consonants in other contexts. Also, certain stressed vowels are shorter in open than in closed syllables. This is of particular interest because it contradicts the hypothesis that speakers tend to keep syllable durations constant.

German. Our study of German duration used the Kiel Corpus of Read Speech, recorded and manually segmented at the Kiel Phonetics Institute and published on CDROM [2].

No German-specific factors or factor levels were found to be necessary. Likewise, the sub-classification scheme followed standard distinctions – the same as used for English [11] – except that vowels were split into central vowels (schwa), diphthongs, and full (non-central) monophthongs.

Among the key findings were that, for consonants in syllable onsets, stops were more than doubled in length by syllabic stress, whereas fricatives were hardly affected. In contrast to English [8], the proportional effect of stress on vowel duration was significantly larger in utterance final position than in utterance medial position.

5. FUTURE DEVELOPMENTS

The percentages of variance accounted for seem to be impressive, in particular in the light of intrinsic within-speaker durational variance [11]. The efficiency by which these systems were constructed is also reason for optimism.

Yet, it is clear that our fundamental understanding of speech timing is limited. The two major areas where progress is needed – discussed next – will require a much deeper understanding of the processes underlying timing.

One area of progress is replacing segmental duration by sub-segmental time-warping, perhaps independently for trajectories of quasi-articulatory parameters that represent concatenative units.

The other area of progress concerns the quality of the input provided by text analysis, not the usage that current duration modules make of this input. Reliably inferring prosodic information from unrestricted text is fundamentally difficult. For example, we have made measurements of timing and F_0 in extremely limited domains (e.g., credit card numbers), and found strong regularities. But these regularities could not easily be explained by general principles. Apparently, certain types of reading materials have to be read in a particular *prosodic style* to sound natural. It may be possible to analyze a given restricted domain into a small number of reading material subtypes, each subtype with its own prosodic style. But we are pessimistic about the applicability of such a style-oriented approach to unrestricted text domains.

6. REFERENCES

1. J. Allen, S. Hunnicut, and D.H. Klatt. *From text to speech: The MITalk System*. Cambridge University Press, Cambridge, U.K., 1987.
2. Institut für Phonetik und digitale Sprachverarbeitung, Universität Kiel. *The Kiel corpus of read speech, vol. 1*, 1994. CDROM.
3. D.H. Klatt. Interaction between two factors that influence vowel duration. *Journal of the Acoustical Society of America*, 54:1102–1104, 1973.
4. A. Maghbouleh. An empirical comparison of automatic decision tree and hand-configured linear models for vowel durations. In *Proceedings of the Second Meeting of the ACL Special Interest Group in Computational Phonology*. Association for Computational Linguistics, 1996.
5. B.M. Möbius and J. P. H. van Santen. Modeling segmental duration in German text-to-speech synthesis. In *Proceedings ICSLP*, pages 2395–2399, Philadelphia, U.S., 1996.
6. C. Shih and B. Ao. Duration Study for the Bell Laboratories Mandarin Text-to-Speech System. In J. P. H. van Santen, R.W. Sproat, J.P. Olive, and J. Hirschberg, editors, *Progress in Speech Synthesis*. Springer-Verlag, New York, 1996.
7. E. Tzoukermann and O. Soumoy. Segmental duration in French text-to-speech synthesis. In *Proceedings of Eurospeech-95*, pages 607–611, Madrid, September 1995.
8. J. P. H. van Santen. Contextual effects on vowel duration. *Speech Communication*, 11:513–546, 1992.
9. J. P. H. van Santen. Deriving text-to-speech durations from natural speech. In G. Bailly and C. Benoit, editors, *Talking Machines: Theories, Models, and Designs*. North Holland, Amsterdam, 1992.
10. J. P. H. van Santen. Analyzing N-way tables with sums-of-products models. *Journal of Mathematical Psychology*, 37(3):327–371, 1993.
11. J. P. H. van Santen. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, 8:95–128, April 1994.
12. J. P. H. van Santen. Combinatorial issues in text-to-speech synthesis. In *Proceedings Eurospeech-97*, Rhodos, Greece, 1997.
13. J. P. H. van Santen and A. L. Buchsbaum. Methods for optimal text selection. In *Proceedings Eurospeech-97*, Rhodos, Greece, 1997.