Efficient Adaptation of TTS Duration Model to New Speakers

Chilin Shih, Wentao Gu, Jan P. H. van Santen*

Bell Labs – Lucent Technologies, Shanghai Jiaotong University*

ABSTRACT

This paper discusses a methodology using a minimal set of sentences to adapt an existing TTS duration model to capture inter-speaker variations. The assumption is that the original duration database contains information of both language-specific and speaker-specific duration characteristics. In training a duration model for a new speaker, only the speaker-specific information needs to be modeled, therefore the size of the training data can be reduced drastically. Results from several experiments are compared and discussed.

1. INTRODUCTION

This paper investigates various methods to adapt an existing text-to-speech (TTS) duration model (the source speaker) to any new speaker (the target speaker). We use Mandarin as a test case. The goal is to capture the target speaker's duration pattern with very few input sentences and to produce a model that performs well on new text input.

A basic assumption is that the source speaker's duration model contains information of both language-specific and speaker-specific characteristics. If we can decompose these two components, then the task of training a target model is reduced to training the speaker-specific component, which in principle can be estimated with less parameters. The size of the training data therefore can be reduced, or if the size remains the same, there would be more observations per parameter, hence a more reliable model.

We will show that if the parameter set is known, the size of the training corpus can be reduced drastically by using a greedy algorithm. In the Mandarin test, 9 greedily selected sentences are sufficient to estimate all 240 parameters from the source model for the target model. The Mandarin model consists of six sub-models, corresponding to six major sound classes. Each sub-model has 14 factors. The 240 parameters represent levels in the 84 factors.

We analyzed the recording of these 9 sentences from six target speakers and evaluated the performance of models fitting different numbers of parameters, ranging from a model that uses only one parameter, the speaking rate, to a model that uses all 240 parameters from the source speaker's model. Both extremes are not ideal. On the one hand, one parameter is not sufficient to capture the target speaker's speaking style. On the other hand, 240 parameters are stretching the limit of the corpus, where many parameters were estimated with just one observation.

We have observed that the ordering of the effects of factor levels tends to be preserved across speakers. So we hypothesize that a new speaker's characteristics can be captured by a set of modification parameters, or weights, which estimate the magnitude of the effects of a factor. This will efficiently reduce the number of parameters to maximally 84 from 240. Our initial investigation shows that this is a viable hypothesis.

2. THE SOURCE MODEL

The model of the source speaker is a Mandarin duration model previously trained on three and half hours of speech from a single speaker [4]. The data are organized in a category tree, and the data in each terminal category are used to train a multiplicative model as in Equation 1 [6, 5]:

$$Dur(p) = Dmean(p) \times D1(f1) \times ... \times Dn(fn) \quad (1)$$

where $Di(fi)$ is a parameter whose value reflects the contribution of factor i when it has level fi, while $Dmean(p)$ denotes the coefficient of the corrected mean duration of the phone p.

The Mandarin category tree is flat, with no more splitting after the initial split by six major sound classes. There are six multiplicative models corresponding to the six major sound classes. Each model was fitted with 14 factors. The sound classes and factors are listed below.

Sound Classes:

1. Vowel (V): 15 vowels, including 4 diphthongs.
2. Fricative (F): 5 fricatives *f, s, x (palatal), S (retroflex), h.*
3. Stop and affricate closure (C): 6 stops and 6 affricates.
4. Stop and affricate burst and aspiration (B): 6 stops and 6 affricates.
5. Nasal coda consonant (N): 2 nasal codas. 1 relatively rare retroflex coda is also included here.
6. Sonorant consonant (S): 8 sonorant/voiced consonants, including 2 nasals, 3 on-glides, *l*, and 2 voiced fricatives.

Factors:

1. Phone identity
2. Tone: Mandarin has 4 lexical tones, one sandhi tone, and a neutral tone which is similar to an unstressed syllable. The tone levels may be combined differently in each terminal category.
3. Preceding phone: Grouped by phone classes. The division is different in each category. This factor has a strong effect in the vowel category, indicating that the vowel duration is affected by the sound class of the preceding phone. This factor has much weaker effect in the initial consonant categories.

4. Preceding tone: Mostly distinguishing whether the preceding tone is a full tone or a neutral tone.
5. Following phone: Grouped by phone classes. The division is different in each category. For example, vowel height distinction has an effect in the fricative category but not in the nasal coda category.
6. Following tone
7. Prominence: Manually transcribed from the speech database.
8. Position of the syllable in the word–distance to the initial position: Typically coding three levels, 0, 1, and 2.
9. Position of the syllable in the word (final)
10. Position of the word in the phrase (initial)
11. Position of the word in the phrase (final)
12. Position of the word in the utterance (initial)
13. Position of the word in the utterance (final)
14. Syllable structure

A total of 292 parameters were used in the original model. Some parameters cannot be estimated from text, such as the prominence level, and some parameters have very little effect, such as *the following tone*, so we trim the parameters to 240 as the basis of corpus text selection.

3. TEXT SELECTION

Once the statistical model is determined and the parameters are known, we can use a model-based greedy algorithm to reduce the corpus size. The best-known algorithm for optimizing coverage is the greedy algorithm as applied to set-covering and matroid-covering problems. For general Analysis-of-Variance models, the optimal text selection is transformed to a linear parameter estimation problem, which can be solved by finding a minimal set of sentences whose design matrix is of full rank [2].

A feature vector $f(x) = (f_1, f_2, ..., f_n)$ corresponding to a given phone segment x can be uniquely represented as a compound row vector $r(f)$, in which each sub-vector $r_k(f)$ encodes the level on the corresponding factor. One way to do this is to have the vector component corresponding to the level in the factor set equal to 1 and the remaining components equal to 0. Usually, the last level is represented as a vector of -1. Then $r(f)$ is defined as $(r_1(f), ..., r_K(f), 1)$, where the last 1 corresponds to the constant term in the model, and K denotes the total number of terms in the Analysis-of-Variance model. The design matrix for a sentence consists of the matrix $X(s)$ whose rows are computed as indicated. The design matrix for a corpus C is a vertical stack of matrices $X(s)$, where s ranges over C. Let $D(C)$ be the corresponding vector of observed duration, P be the column vector of parameters, then we have $D(C) = X(C) \cdot P$. P is estimable if and only if the matrix $X(C)$ is of full rank. The optimization problem can now be formulated as finding a minimal subset of C, say, C', so that $X(C')$ is also of full rank. We can achieve further reduction of text selection if the parameters in multiple models are considered simultaneously. This is done by concatenating all the parameters from different models in the vector P. This kind of concatenated design matrix can handle any kind of classification tree with each leaf characterized by an Analysis-of-Variance model.

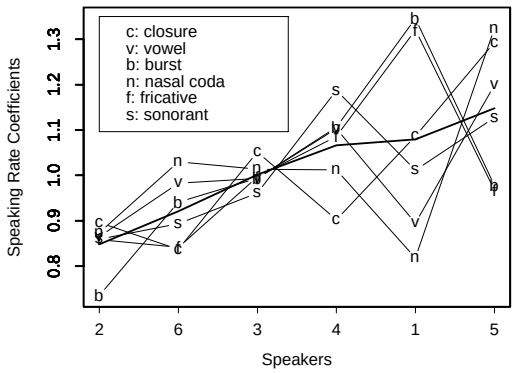

Figure 1: Speaking rate coefficients of different sound classes from six speakers

4. DATA

Six sentences were selected from a corpus of 15620 newspaper sentences by the first run of the greedy algorithm selection. These sentences can be used to estimate the 240 parameters we targeted in the greedy search. To increase the number of observations we run the greedy algorithm iteratively to select a total of four sets of sentences, each set containing 6 to 7 sentences. The sentences were recorded by 10 speakers; some read it with varied speaking rate. We have analyzed the first run of sentences in normal speaking rate from six of the speakers. After the initial analysis, we found that some parameters cannot be estimated due to reading discrepancies that destroy some factor levels, primarily from the insertion of glottal stops before word-initial low vowels. Three more sentences were selected to compensate for the loss.

The analyses performed for this paper are based on this 9-sentence, 6-speaker database. It contains 4346 phones, with 1529 vowels, 599 closures (phrase-initial ones excluded), 662 bursts with aspiration, 561 nasal codas, 401 fricatives, and 594 sonorants.

We use this corpus to test a few models, starting with one that estimates only one parameter per target speaker, the mean speaking rate.

5. SPEAKING RATE

Some voice conversion systems such as the one reported in [1] simply convert the source duration model to the new speaker by detecting and changing the speaking rate. Speaking rate can be estimated quickly given any input, which is an undeniable advantage [7]. However, its effectiveness rests on whether the duration of phones is stretched or compressed uniformly when speaking rate changes [3]. We compare two models: using only the coefficients of the source model described in Section 2 to predict the duration of the six new speakers, or to apply a speaking rate multiplier for each speaker to the original model. We compare these two conditions using root mean squared deviation of observed and predicted duration of the 36 models (6 speakers, each with 6 models for different sound classes). 14 models perform better with the speaking rate multiplier, 19 models perform worse, and 3 models are practically identical. One question is why the rate multiplier does not work as well as expected.

We investigate this question further by training the usual six models for each speaker, but using only two factors: phone identity and speaker identity. The coefficients obtained from the speaker factor are plotted on the y-axis in Figure 1. The six speakers, ordered from fast (with smaller coefficients) to slow (with larger coefficients), are shown on the x-axis. The plotting symbols indicate the models by sound classes: v, c, b, n, f, s represent vowel, stop and affricate closure, stop and affricate burst, nasal coda, fricative, and sonorant, respectively. The thick line going through the middle shows the mean rate of each speaker. If phone durations are stretched uniformly from fast to slow speakers, we expect the lines connecting phone classes not to cross. This is apparently not the case.

The total lack of consistency in the scaling of sound classes across speakers is interesting, particularly considering that two pairs of co-varying coefficients B/F and V/N, which show that there is a high level of consistency within speakers. For all speakers, the burst and fricative coefficients are quite similar, with four speakers showing nearly identical values. The vowel and nasal coda coefficients of each speaker are also comparable. Note that fricatives and plosive bursts are similar, in particular, affricates may be analyzed as the combination of a stop followed by a fricative. Vowels and nasal codas are similar in that they comprise the rhyme of the syllable. Apparently the shared properties of B/F or V/N are subject to contextual factors in the same way for the same speaker, which contributes to the perceived speaking style of the speaker. Across speakers the scaling of sound classes becomes unpredicatable. A good example is that bursts and fricatives are proportionally long for speaker 1 but are short for speaker 5. At least part of the durational characteristics of a speaker is revealed in the length proportion of sound classes. Modifying speaking rate uniformly will not be effective.

If we use pre-selected input text, then we are able to estimate many more parameters given that the text is optimized with reference to the intended parameters. One additional advantage of using pre-selected sentences is that many procedures such as speech segmentation and parameter identification can be automated. In the following section we identify a set of parameters that can be used effectively to adapt duration models to the target speaker.

6. THE TARGET MODEL

Given that our corpus was based on greedily selected sentences and was amended for performance errors, we can use the corpus to estimate all the parameters as in the model of the source speaker, following Equation 1. The performance of these models, referred to collectively as Model I, is used as the baseline performance guide to judge subsequent experiments. Model I is trained separately for six speakers, where the tree categorization and factor levels were identical to the final revision of the model with 240 parameters described in Section 2. Table 1 gives the correlation scores of the predicted duration and the observed duration from 36 models, sound classes in rows and speakers in columns.

We proceed to test Model II, given in Equation 2, which assumes common coefficients $Dmean(p)$ and $Di(fi)$ for all speakers. For each speaker K we only need to estimate a set of modification parameters ki.

Speaker	1	2	3	4	5	6
Vowel	0.83	0.77	0.77	0.80	0.83	0.81
Burst/Asp	0.94	0.93	0.93	0.94	0.95	0.91
Closure	0.77	0.73	0.80	0.75	0.71	0.78
Nasal Coda	0.75	0.71	0.65	0.81	0.71	0.74
Fricative	0.87	0.78	0.75	0.81	0.70	0.73
Sonorant	0.82	0.66	0.74	0.71	0.80	0.70

Table 1: Correlation of estimated and observed duration–Model I

$$Dur(p)_K = Dmean(p)^k \times D1(f1)^{k1} \times ... \times Dn(fn)^{kn} \quad (2)$$

Equation 2 makes sense if the parameter coefficients of a given factor from different speakers are in scale. It was shown that vowels in English do indeed maintain their scale under different prosodic conditions [5]. Our data show that the phone levels in the phone identity factors are typically in scale across speakers. We show the cross-speaker coefficients of the vowel identity factor and the tone factor in Figures 2 and 3. In Figure 2, the apical vowels *J, Q* are consistently the shortest for all speakers, high vowels *i, u, U* are in the next group, while low vowels *a* and diphthongs are the longest.

Figure 3 illustrates the cross-speaker consistency from the tone factor of the vowel category. which shows the effect of tones on vowel duration. The factor successfully captures an impression noted during the recording session that some speakers maintain clear contrast between full tone and neutral tone syllables, while some hardly make any distinction. The tone coefficients show that for all speakers, the neutral tone labeled as *0* has the strongest effect in shortening vowels. The magnitude of this effect is different for each speaker. It appears that the target speaker's characteristics can be captured by multiplying the source speaker's coefficients with a weight. In this case, only one weight, or modification parameter, needs to be estimated instead of 5 tone parameters.

We proceed with experiment 1 to assess the validity of this assumption for all factors. Given a matrix A of $m \times n$ dimensions containing coefficients from a factor with m levels and n speakers, we want to know whether A can be approximated by $F \cdot W$, where F is a $m \times 1$ vector functioning as the common parameter vector of factor f for all speakers, and W is a $1 \times n$ vector of weights. We obtained F and W by singular value decomposition, which returns two orthogonal matrices and a diagonal matrix $A = UDV^t$, where D is the diagonal matrix. The best approxi-

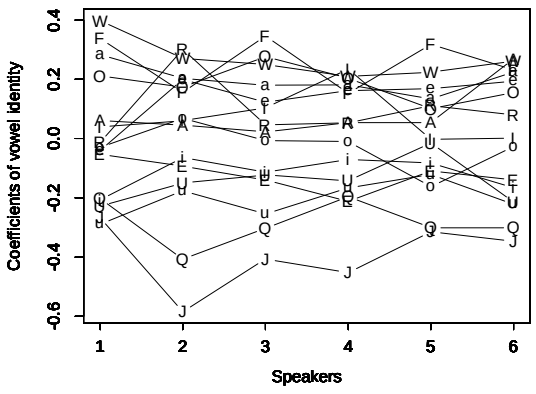

Figure 2: The coefficients of the vowel identity factor

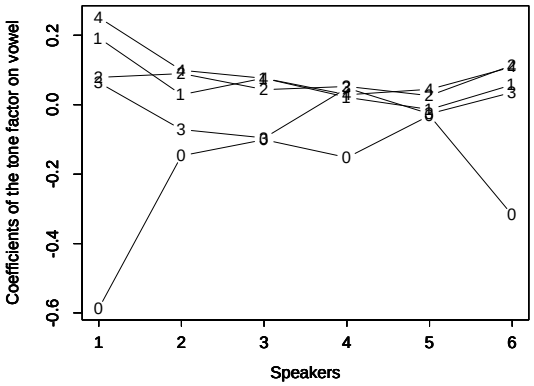

Figure 3: The effects of tones on vowel duration.

mation to A is $\sqrt{d_1} \cdot u_1 \cdot v_1$. We take u_1, the first column vector of U, as the common parameter vector F, and $\sqrt{d_1} \cdot v_1$ as the weight vector W, where v_1 is the first row vector of V^t. We next compute eigenvalues. The first eigenvalue is very high in the majority of matrices, suggesting that most of the variation can be captured in the first eigenvector and the assumption underlying Equation 2 is supported.

There is an alternative method to calculate W. We substituted the factor level codes in the duration data matrix of each speaker with the corresponding F entries and fitted a robust regression model to predict the observed duration. The coefficients of this model are used as weights W'. We then estimate duration with F and W'. The correlation scores of the predicted and observed duration are given in Table 2. The result is comparable to the baseline correlation scores shown in Table 1, which is very good considering the 3 to 1 reduction of the number of parameters.

Experiment 1 was performed primarily to test the consistency across speakers and to assess the possibility of using weights. The common vector F was estimated with the target speaker's data, and all speakers were reading the same sentences. The results in Table 2 were obtained from a test on the training data. We continued with experiment 2 to see how the described method performs with testing data.

In experiment 2, the factor level codes in the data matrix were replaced with corresponding coefficients from the models of the source speaker, rather than with F. Again, we fit a robust regression model to predict the target speaker's duration, and use the coefficients from this model as weights, adapting the source model to the target model. The correlation scores of the observed and predicted duration are given in Table 3. With the exception of closure duration, the results from all other sound classes are fairly good. In a few cells the scores are even better than those from experiment 1. Even though the correlation scores from experiment 2 are in general worse than those from experiment 1, the models of experiment 2 should perform better on new sentences since the source speaker's model is estimated from a much larger corpus, and is therefore less affected by random variations in speech and the peculiarity of individual sentences.

7. CONCLUSION

We have shown that a target speaker's duration model can be effectively adapted from a source speaker's model. This is achieved

Speaker	1	2	3	4	5	6
Vowel	0.81	0.74	0.76	0.78	0.79	0.80
Burst/Asp	0.91	0.93	0.91	0.93	0.94	0.91
Closure	0.73	0.73	0.79	0.69	0.66	0.78
Nasal Coda	0.70	0.67	0.61	0.76	0.59	0.72
Fricative	0.83	0.75	0.71	0.80	0.65	0.69
Sonorant	0.67	0.54	0.68	0.56	0.73	0.69

Table 2: Correlation of estimated and observed duration–Model II, experiment 1, using weights and common vector obtained from singular vector decomposition to predict target speaker's duration.

Speaker	1	2	3	4	5	6
Vowel	0.74	0.67	0.72	0.75	0.77	0.77
Burst/Asp	0.89	0.93	0.89	0.94	0.91	0.90
Closure	0.61	0.61	0.61	0.67	0.50	0.39
Nasal Coda	0.65	0.69	0.55	0.75	0.63	0.71
Fricative	0.82	0.76	0.68	0.80	0.64	0.65
Sonorant	0.75	0.47	0.58	0.55	0.72	0.65

Table 3: Correlation of estimated and observed duration–Model II, experiment 2, using weights and the source speaker's coefficients to predict target speaker's duration.

in two stages: text selection and weight estimation. Assuming the source speaker's model, the size of the training corpus for the target speaker can be reduced dramatically (nine sentences for the present study). The small database reduces data collection and processing time. The source speaker's model can be adapted to the target speaker with a set of weights, which is an effective way to capture speaker-specific characteristics.

8. REFERENCES

1. Arslan, L. M., and Talkin, D. Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum. In *EuroSpeech'97* (Rhodes, Greece, 1997), vol. 3, pp. 1347–1350.
2. Buchsbaum, A. L., and van Santen, J. P. H. Selecting training text via greedy matroid covering. Tech. memo, AT&T Bell Laboratories, 1994.
3. Klatt, D. H. Interaction between two factors that influence vowel duration. *JASA 54*, 4 1973, 1102–1104.
4. Shih, C., and Ao, B. Duration study for the Bell Laboratories Mandarin text-to-speech system. In *Progress in Speech Synthesis*, J. van Santen, R. Sproat, J. Olive, and J. Hirschberg, Eds. Springer, New York, 1997.
5. van Santen, J. P. H. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language 8*, 2 1994, 95–128.
6. van Santen, J. P. H., and Olive, J. P. The analysis of contextual effects on segmental duration. *Computer Speech and Language 4* 1990, 359–391.
7. Verhasselt, J. P., and Martens, J.-P. A fast and reliable rate of speech detector. In *ICLSP 96* (Philadelphia, 1996), vol. 4, pp. 2258–2261.