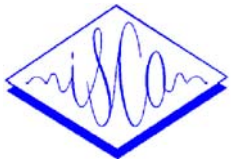

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

AN EFFICIENT SPEAKER ADAPTATION METHOD FOR TTS DURATION MODEL

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

Shanghai Jiao Tong University, Shanghai, China*
Bell Labs-Lucent Technologies, Murray Hill, U.S.A.

ABSTRACT

This paper is a continuation of our previous study [1], where an efficient speaker adaptation method was proposed for TTS duration model. The goal was achieved by text selection and weight estimation. The result there was preliminary because it's only derived from one sentence set. After the analysis on multiple sentence sets, we can now evaluate the robustness of the method better and hence a more confident conclusion is given. Based on the observation that some language-specific information is well preserved across speakers, the proposed method is supported. By a further comparison between various adaptation models, the linear weighted model shows the best performance, and therefore presents an efficient way to adapt the duration model from the source speaker to target speakers with a very small training corpus.

1. INTRODUCTION

Personalization is an important demand for future TTS systems, where speaker-adaptive prosody modeling is a critical technology. Study on this issue is receiving more attention today [2]. Considering the fact that in practical applications only a realistically very small corpus can be employed for training a new speaker, the adaptation task is more challenging.

One possible solution to this problem is, assuming a general model formula, only the coefficients of the model need modification to feature a specific speaker, therefore the adaptation task will be drastically simplified, and the training corpus will be decreased a lot. In [1], we've proposed a method to adapt the source speaker's duration model to target speakers. Two steps were involved: text selection and weight estimation. The number of parameters was greatly reduced by introducing a linear weighted model. The result was quite inspiring. However, the conclusion was preliminary because it's derived from only one sentence set. This paper reports our further study on this issue. With three sentence sets available, we can compare the performance of various models more reliably.

Based on the observation that duration coefficients tend to show consistent pattern with varied scales across speakers, the linear weighted model is supported, where only a set of modification weights need to be adapted for the target speaker. Therefore the size of training data can in principle be further decreased, or if the corpus size remains the same, a more reliable model can be derived. By dividing multiple sentence sets separately into the training data and testing data, we verify that this model is less liable to sentence effect, and hence more robust for generalization.

We next compare 4 different models directly adapted from the source speaker to target speakers, with the number of adapting parameters increasing successively. Both extremes fail to capture the speaker styles, while the linear weighted model is a best balance among various models, achieving good speaker-adaptation performance with a quite small training database.

2. THE SOURCE MODEL

The source duration model is the one previously trained on a single speaker's speech based on a total of 424 Mandarin sentences [3], and established by using the statistical method proposed by van Santen [4]. The model consists of 6 categories: (1) vowels; (2) plosive closure; (3) plosive burst and aspiration; (4) nasal codas; (5) fricatives; (6) sonorant consonants. Within each category, a multiplicative model is established for segmental duration [4]:

$$Dur(f) = Dmean \times D_1(f_1) \times \cdots \times D_n(f_n) \quad (1)$$

where $D_i(f_i)$ is the parameter whose value reflects the contribution of the ith factor when it takes the level f_i, while *Dmean* is the corrected mean duration for the phone category. As shown in [1], altogether 14 factors are used in the modeling.

In our Mandarin corpus, each sentence is composed of several phrases, therefore each sentence functions as an utterance. Beside punctuation marks, all spontaneously inserted pauses in the utterance are also coded as phrase boundary. We adopt this coding method because many previous studies have revealed that pre-pausal lengthening is much more noticeable than phrase-final lengthening. Of course, during the text selection phase, phrasing is coded solely on the basis of punctuation.

Each factor may have 2 to 15 levels, and for the same factor the number of levels may vary across phone categories. We slightly modify the factor levels for text selection. Some more levels are added to give more flexibility to the modeling for new speakers, while any factor whose levels cannot be reliably predicted from text is excluded, such as degree of prominence.

We temporarily assume that the formula of the source model can be generalized to new Mandarin speakers. However, in the later section, this assumption will be verified by the good fitting performance of the model on all the new speakers.

3. TEXT SELECTION AND DATA COLLECTION

Since the multiplicative model is a special case of the analysis-of-variance model, the optimal text selection can be accomplished by a model-based greedy algorithm [3]. We further modify the algorithm for multi-model cases by considering all categories simultaneously [1], such that the selected sentences can be further reduced. We run the algorithm iteratively on a corpus of 15,620 newspaper

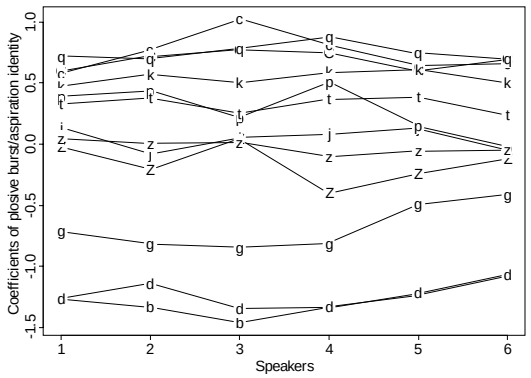

Figure 1: The log coefficients of plosive burst/asp identity

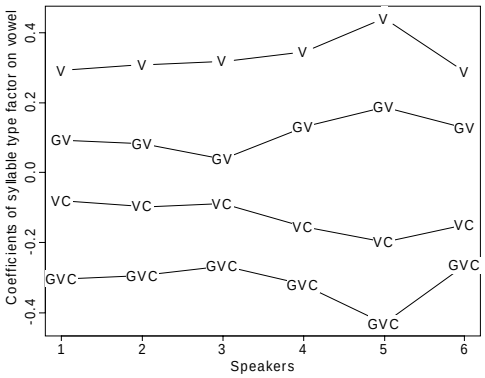

Figure 2: The log coefficients of syllable type factor on vowel. V: open syllable without glide; GV: open syllable with glide; VC: closed syllable without glide; GVC: closed syllable with glide.

sentences, from which 3 sentence sets are finally selected, each containing 6 sentences to achieve a full covering.

All the 18 sentences are recorded for multiple Mandarin speakers, some with varied speaking rates. In our current study, only the data for 6 speakers in normal speaking rate have been used. The recording procedure was carefully controlled to make the speaking rate consistent within one speaker, and for each sentence the recording will be repeated until it is read fluently. All the recorded speech is then manually labeled. Since all the phrase-initial (or post-pausal) plosive closures are merged with silence, they are excluded from the database.

The 6-speaker database on which this study is based consists of a total of 12,781 phone segments, including 4390 vowels, 2032 plosive burst and aspiration, 1823 plosive closure, 1664 nasal codas, 1114 fricatives, and 1758 sonorant consonants. For reference, using the slightly modified factor levels, we also re-estimate the model for the source speaker based on a sub-database of totally 3669 segments, whose corresponding corpus is different from the sentence sets we have just selected.

4. LINEAR WEIGHTED MODEL

Assuming that equation (1) gives a general formula applied to all Mandarin speakers, we estimate the model parameters for the 6 speakers each with 6 categories respectively. Different from [1], now the models are trained on 3 sentence sets, and therefore should be more robust than before. The correlation coefficients between the observed and estimated duration are given in Table 1, where the bottom row represents the overall correlation scores on all phone classes. The re-estimated source model is also presented here as speaker 0. Although the model was originally derived from the source speaker, it can be generalized to other speakers very well, and actually it fits a little better on all target speakers than on the source speaker.

From the comparison of all the models, we find that for most

Speaker	0	1	2	3	4	5	6
vowel	0.73	0.81	0.75	0.73	0.75	0.81	0.77
burst/asp	0.91	0.92	0.90	0.90	0.92	0.93	0.88
closure	0.67	0.66	0.67	0.74	0.69	0.66	0.65
nasal coda	0.49	0.61	0.57	0.52	0.69	0.61	0.61
fricative	0.66	0.76	0.72	0.73	0.72	0.61	0.66
sonorant	0.59	0.78	0.66	0.69	0.65	0.69	0.70
Total	0.85	0.87	0.89	0.86	0.88	0.91	0.88

Table 1: Correlation scores for multiplicative models

duration factors, the coefficients show quite consistent patterns across speakers. In [1], we have shown the cross-speaker coefficients for two factors: vowel identity and tone identity of vowel, both of which exhibit better consistency in current study involving multiple sentence sets. In this paper, we'll present more evidence on other two factors: the identity of plosive burst/aspiration, and syllable type factor of vowel.

Figure 1 plots the log coefficients of the identity of plosive burst and aspiration. The ordering of the 12 phones is preserved quite well across speakers. Consistently aspirated affricates are the longest, the next is aspirated stops (within which the velar *k* is consistently longer), and the next unaspirated affricates (within which a fairly consistent tendency is observed: palatal *j* > dental *z* > retroflex *Z*), while unaspirated stops are the shortest (within which the velar *g* is the longest, and the dental *d* is a little longer than the labial *b*). This ordering coincides with previous findings on Mandarin consonant duration, revealing the strong intrinsic effect on duration of different manner of articulation.

In Figure 2 the log coefficients of the syllable type factor on vowel category are plotted. Here syllable structures are grouped into 4 levels, according to the presence or absence of glide and nasal coda. A very consistent tendency is observed: open syllable without glide > open syllable with glide > closed syllable without glide > closed syllable with glide. This exactly coincides with the finding in [3], and reflects a kind of compensatory effect to keep the duration of the whole syllable constant, i.e., the more complicated the syllable structure is, the shorter the duration of the vowel within it tends to be.

In both figures, the coefficients for different speakers are usually in different scales. For example, in Figure 2 the coefficients for speaker 5 show the largest scale, indicating that this speaker has stronger syllable compensatory effect than others. The different scales of effect, may well characterize the speaker style. Assuming this, only one scale parameter per factor is needed to model a specific speaker, and hence the model can be drastically simplified by the following equation:

$$Dur(f) = Dmean \times D_1(f_1)^{k_1} \times \ldots \times D_n(f_n)^{k_n} \qquad (2)$$

where $D_i(f_i)$'s are common parameters across speakers which are known a priori, and only *Dmean* and a set of weights k_i need re-estimation for the specific speaker. We name equation (2) as linear weighted model by noting its form in log domain.

Speaker	1	2	3	4	5	6
vowel	0.81	0.74	0.72	0.73	0.80	0.78
Burst/asp	0.90	0.90	0.90	0.92	0.92	0.88
closure	0.63	0.66	0.72	0.68	0.64	0.64
nasal coda	0.57	0.56	0.52	0.69	0.59	0.59
fricative	0.76	0.72	0.72	0.72	0.59	0.67
sonorant	0.78	0.64	0.69	0.65	0.67	0.69
Total	0.87	0.89	0.86	0.87	0.90	0.88

Table 2: Correlation scores for linear weighted models

We assess the validity of the assumption by the following method. Given a matrix A of $m \times n$ dimensions, containing coefficients from a factor f with m levels and n speakers, we want to know whether A can be well approximated by $F \cdot W$, where F is an $m \times 1$ vector functioning as the common parameter vector of factor f for all speakers, and W is a $1 \times n$ vector of weights. We obtain F and W by singular value decomposition, which returns $A = UDV^t$, where U and V are both orthogonal matrices, while D is a diagonal matrix. The best one-rank approximation to A is $\sqrt{d_1} \cdot u_1 \cdot {v_1}^t$, where d_1 is the maximal eigenvalue of A^tA. We take u_1, the first column vector of U, as the common parameter vector F, and $\sqrt{d_1} \cdot {v_1}^t$ as the weight vector W, where ${v_1}^t$ is the first row vector of V^t.

How good this approximation is, can be evaluated by the ratio of d_1 to other eigenvalues. For about four fifth of the factors containing more than 2 levels, the value of d_1 accounts for more than 85% of the sum of all eigenvalues, suggesting that most of the variations can be captured by the first eigenvector. This gives powerful grounds for the scale relationship.

We employ an alternative way to calculate the weights directly from the data. Using the common parameter vector F derived from the above analysis, we can fit equation (2) by a robust linear regression model. The coefficients obtained this way are used as weights W', and the duration is then estimated with F and W'. The correlation scores of the estimated and observed duration are given in Table 2. By comparison with Table 1, the difference between the performance of the two models is negligible, hence the assumption of equation (2) is supported.

However, similar to [1], in the above analysis we didn't separate the training data from the testing data. The performance is tested on the same database from which the model parameters are just derived. This analysis is fragile, because there's no way to know whether it performs well on new text input. With multiple sentence sets available now, we can investigate the variation of the models across different training data. We run both models on each of the 3 sentence sets respectively, and big variances among the cross-set coefficients are found, though the overall tendencies are preserved. This is not surprising, because the 6-sentence corpus is too small to resist the natural variability.

Since our adaptation is based on minimized sentence sets, which is the most important advantage of our method, the generalization ability is critical. This is strongly supported by the sentence effect we've just observed. Therefore, we design experiment 1 to evaluate the robustness of the two models.

In experiment 1, we use the first sentence set as the training data from which the parameters of both models are derived, while the other two sentence sets are used as the testing data. For the linear weighted model , we still use the common parameter F derived from all the 3 sentence sets, assuming this is a priori knowledge. The performance of the two models on both one training set and two testing sets are given in the left and middle columns of Table 3 respectively. For concision, here only the overall correlation scores of the estimated and observed duration on all phone categories are shown.

Speaker	Original model			LW model-Exp1			LW model-Exp2		
	Set 1	Set 2	Set 3	Set 1	Set 2	Set 3	Set 1	Set 2	Set 3
1	0.90	0.80	0.79	0.89	0.84	0.84	0.88	0.81	0.78
2	0.92	0.79	0.83	0.91	0.85	0.86	0.90	0.83	0.85
3	0.88	0.78	0.80	0.87	0.82	0.84	0.85	0.80	0.83
4	0.89	0.79	0.77	0.86	0.86	0.83	0.85	0.85	0.82
5	0.92	0.86	0.86	0.91	0.88	0.88	0.90	0.87	0.88
6	0.90	0.79	0.79	0.89	0.84	0.84	0.88	0.82	0.81

Table 3: Correlation scores on both the training data (set 1) and the testing data (set 2 and 3)

Although on the training data the performance of the linear weighted model is a little inferior to (but approaches very well) that of the original multiplicative model, it obviously performs better than multiplicative model on the testing data. In fact this is reasonable, because estimating all the parameters are stretching the limit of the corpus, where many parameters are estimated with only very few observations. This experiment gives the confidence that the linear weighted model is less liable to sentence effect and hence more reliable.

5. COMPARISON OF VARIOUS ADAPTATION MODELS

To further investigate the validity of the linear weighted model, we'll compare the performance of four different adaptation models, with the number of adapting parameters increasing successively: (1) The simplest model, with only one adapting parameter (mean speaking rate) per speaker; (2) Speaking rate multiplier is estimated for each phone category respectively; (3) The linear weighted model, with one weight per factor in addition to one rate multiplier per category; (4) Independently re-estimate all parameters in the original multiplicative models.

In the above analysis, the common parameter F was derived from the target speakers' data, and all target speakers read the same 18 sentences. This is not helpful for scientific evaluation. In the following design, we'll directly adapt from the source speaker, using the coefficients of the source model as the basis of adaptation instead of F. In this way, the various models are comparable. Table 4 gives the median percentage error (defined as the median of the ratio of absolute estimation error to observed duration) between the estimated and observed duration for all the four adaptation models. Here Model 0 stands for directly applying the model of the source speaker to new speakers without any adaptation, which only gives a baseline for comparison. The rightmost column of Table 4 gives the number of *free* parameters estimated in each model.

Speaker	1	2	3	4	5	6	Par. number
Model 0	31.3	30.2	26.4	23.2	19.3	27.1	0
Model 1	30.7	23.7	26.1	25.0	19.4	24.2	1
Model 2	27.5	23.4	25.7	23.4	18.5	24.2	6
Model 3	22.3	20.0	21.4	21.1	16.3	20.2	75
Model 4	19.2	19.1	20.9	20.4	15.6	19.2	178

Table 4: Median percentage error (%) of various models

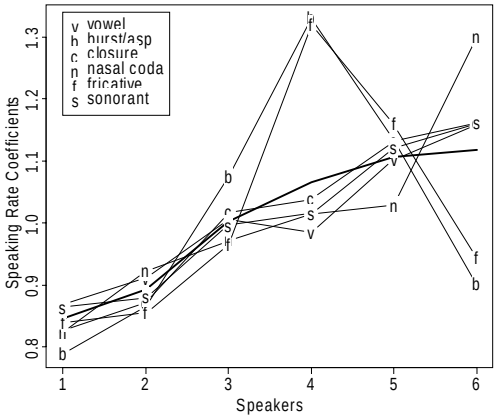

Figure 3: Speaking rate coefficients of different phone classes. The actual speaker number, from left to right, is 2, 6, 3, 1, 4, 5.

The simplest and quickest adaptation method (Model 1), as used by some voice conversion systems [6], is only adapting the mean speaking rate for a new speaker. However, from Table 4, we find that among the 6 speakers, only for 2 speakers Model 1 performs better than Model 0, for 3 speakers their performance are almost identical, and for 1 speaker Model 1 performs even worse than Model 0.

The failure of Model 1 is due to its over simplicity. To further investigate this problem, we calculate the speaking rates for each phone category respectively. Different from Figure 1 in [1], the rate coefficients in Figure 3 are obtained by simply normalizing among speakers the average segmental duration of each phone category. Still the 6 speakers are ordered from fast to slow, and the middle thick line represents the mean rate of each speaker. Although the rate coefficients concentrate around the mean rate line better than in [1], Figure 3 still leads to the same conclusion as before, particularly noting the big scale deviation of the co-varied pair b/f from other phone categories.

Aware of the rate inconsistency across phone categories, we proceed to estimate the rate multiplier (the coefficient *Dmean* in equation (1)) for each phone category respectively. As expected, by adding 5 more parameters Model 2 performs consistently better than Model 1, however, the improvement is quite limited. It's interesting to note that the results in Table 4 l coincide well with Figure 3. The 3 speakers (1, 4, 5) showing big rate inconsistency across phone classes all achieve some improvements in the evolution from Model 1 to Model 2, while the other 3 speakers just show the opposite. Therefore, Model 2 is still too simple to capture the necessary features of a specific speaker. We conjecture that even within the same phone class, the duration doesn't stretch uniformly across speakers. Actually this has been evidenced by the fact that the coefficients of a factor usually show different scales across speakers.

Considering these cross-speaker feature variations, Model 3 has been proposed, where a set of weights are estimated in addition to *Dmean*. In Table 4, it's noticeable that the biggest improvement takes place after Model 3 is introduced. It's not strange that its performance is not as good as that of Model 4, which re-estimate all the parameters in the multiplicative model to achieve a best fit. However, as we have discussed in the previous section, the better performance of Model 4 is due to over-estimation, and will be fragile in generalization.

Similar to the analysis in section 4, we design experiment 2 to evaluate the generalization ability of Model 3 by separating the training data (set 1) from the testing data (set 2 and 3). The correlation scores of the estimated and observed duration are given in the right column of Table 3. Resembling the result of experiment 1, Model 3 fits better on the two testing sets than Model 4, indicating its better robustness again. Considering that in this experiment the model is adapted from a random speaker whose source model is trained on a quite different corpus, its generalization ability is even more inspiring.

With the number of parameters increasing successively from 0 (no adaptation) to 174 (full re-estimation), the performance of the models in Table 4 is improved step by step. Having too few parameters, Model 1 and 2 fail to capture great speaker style variations. Model 4 should perform best if a large corpus were available, which however is not the goal for our adaptation task. Considering the fact in Table 4 that Model 3 approaches Model 4 very well with only 42% parameter cost of the latter, Model 3 is the best choice to achieve good adaptation on small training corpus, which is further confirmed by experiment 2.

6. CONCLUSION

Based on the analysis of multiple sentence sets, we can now draw more confident conclusion than in [1]. Some language-specific information has been shown very helpful in speaker-adaptive duration modeling. First, the formula of the source speaker's model is successfully shared by new speakers, which gives the grounds for pre-selecting a minimal training corpus, and hence reduce the data collection and processing time greatly. Second, the duration coefficients show quite consistent patterns across speakers, based on which the linear weighted model is proposed, therefore further decrease the parameters to a set of weights. By comparison of various adaptation methods, the linear weighted model proves an efficient way to capture the target speaker's duration pattern with very few sentences, and performs quite robustly on new text input.

There are still a lot of margins left for future study. For example, the basis of our adaptation method is multiplicative model, which is not definitely the best choice for the source duration model. How to generalize our method to an arbitrary sums-of-products model [4], needs to be further addressed.

7. REFERENCES

1. Shih, C., Gu, W., and van Santen J.P.H. Efficient adaptation of TTS duration model to new speakers. In *ICSLP 98* (Sydney, Australia, 1998), vol.1, pp.177-180.
2. López-Gonzalo, E., Rodríguez-García, J.M, Hernández-Gómez, L., and Villar, J.M. Automatic prosody modeling for speaker and task adaptation in text-to-speech. In *ICASSP 97* (Munich, Germany, 1997), vol.2, pp.927-930.
3. 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.
4. van Santen, J.P.H. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and language 8*, 1994, 95-128.
5. van Santen, J.P.H, and Buchsbaum, A.L. Methods for optimal text selection. In *EuroSpeech 97* (Rhodes, Greece, 1997), vol.2, pp.553-556.
6. 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.