Phone Duration Modeling Using Gradient Tree Boosting

Junichi Yamagishi, Hisashi Kawai, and Takao Kobayashi. Phone duration modeling using gradient tree boosting. Speech Communication, 50(5):405-415, May 2008.

Phone Duration Modeling Using Gradient Tree Boosting

Junichi Yamagishi[a,b*] Hisashi Kawai[c,a] and Takao Kobayashi[b]

[a]Spoken Language Communication Research Laboratories,
Advanced Telecommunications Research Institute International,
2-2-2, Hikaridai, Seika, Soraku, Kyoto, 619-0288, Japan

[b] Interdisciplinary Graduate School of Science and Engineering,
Tokyo Institute of Technology,
4259-G2-4 Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa, 226-8502, Japan

[c]KDDI R&D Laboratories, 2-1-15, Ohara, Fujimino, Saitama, 356-8502, Japan

In text-to-speech synthesis systems, phone duration influences the quality and naturalness of synthetic speech. In this study, we incorporate an ensemble learning technique called *gradient tree boosting* into phone duration modeling as an alternative to the conventional approach using regression trees, and objectively evaluate the prediction accuracy of Japanese, Mandarin, and English phone duration. The gradient tree boosting algorithm is a meta algorithm of regression trees: it iteratively builds the regression tree from the residuals and outputs weighting sum of the regression trees. Our evaluation results show that compared to the regression trees or other techniques related to the regression trees, the gradient tree boosting algorithm can substantially and robustly improve the predictive accuracy of the phone duration regardless of languages, speakers, or domains.

1. Introduction

In text-to-speech synthesis, phone duration determines the rhythm and tempo of synthetic speech, and thus influences quality and naturalness. In general, control of phone duration can be viewed as a problem of estimating nonlinear prediction functions using several phonetic, prosodic, and linguistic features obtained from input text as explanatory variables. From this point of view, many researchers have developed effective methods for phone duration modeling using, e.g., linear regression [1], regression trees [2], neural networks [3],[4], and so on. However, these methods are not always satisfactory for reproducing phone duration, as shown in [5].

In this study, we incorporate a promising approach called *gradient tree boosting* (GTB) [6],[7] into phone duration modeling as an alternative to the conventional approach using regression trees. The GTB algorithm is a meta algorithm for constructing multiple regression trees and taking advantage of them. The algorithm iteratively utilizes the residuals of the current prediction function as the training data for the next regression tree to

*The author presently belongs to University of Edinburgh.

be constructed, and builds an augmented predictive function by simply combining the iteratively constructed regression trees. Without modifying building algorithms of each regression tree, it surprisingly increases the expressive ability of the regression trees much more than with a single large regression tree [8]. To confirm the effectiveness of this approach and evaluate its language dependency, we have applied it to Japanese, Mandarin, and English phone duration modeling and have objectively compared the approach with conventional phone duration modeling approaches.

This paper is organized as follows. Section 2 gives an overview of the GTB algorithm for reference. Experimental conditions and the objective evaluation results of Japanese, Mandarin, and English phone duration modeling are described in Section 3. Comparison results with several conventional duration modeling techniques are also described in this section. Section 4 summarizes our findings.

2. Gradient Tree Boosting

In this section, we review gradient tree boosting (GTB) [6],[7]. We define explanatory variables and a target value as $\boldsymbol{x} = (x_1, x_2, \cdots, x_K)$ and y, respectively. Let $\{y_i, \boldsymbol{x}_i\}_1^N$ be a set of training data including N pairs. The GTB algorithm iteratively constructs M different regression trees $h(\boldsymbol{x}, \boldsymbol{a}_1) \ldots h(\boldsymbol{x}, \boldsymbol{a}_M)$ from the set of training data and constructs the following additive function $F(\boldsymbol{x})$

$$F(\boldsymbol{x}) = \beta_0 + \sum_{m=1}^{M} \beta_m \, h(\boldsymbol{x}, \boldsymbol{a}_m), \tag{1}$$

where β_m and $\boldsymbol{a}_m$ are a weight and vector of parameters for the the m-th regression tree $h(\boldsymbol{x}, \boldsymbol{a}_m)$, and β_0 is an initial value. Both the weight β_m and the parameters $\boldsymbol{a}_m$ are iteratively determined from $m = 1$ to $m = M$ so that a loss function $\Psi(y, F(\boldsymbol{x}))$ is minimized. Now, we define an additive function that is combined from the first regression tree to the $(m-1)$-th regression tree as $F_{m-1}(\boldsymbol{x})$. The weight β_m and the parameters $\boldsymbol{a}_m$ for the m-th regression tree should be determined as follows:

$$(\beta_m, \boldsymbol{a}_m) = \underset{\beta, \boldsymbol{a}}{\operatorname{argmin}} \sum_{i=1}^{N} \Psi(y_i, F_{m-1}(\boldsymbol{x}_i) + \beta h(\boldsymbol{x}_i, \boldsymbol{a})), \tag{2}$$

where $F_0(\boldsymbol{x})$ is an initial value and given by $F_0(\boldsymbol{x}) = \beta_0 = \operatorname{argmin}_\beta \sum_{i=1}^{N} \Psi(y_i, \beta)$.

However, in general, it is not straightforward to solve Eq. (2). Therefore, gradient tree boosting separately and approximately estimates $(\beta_m, \boldsymbol{a}_m)$ with a simple two-step procedure [6]. In the estimation of the parameters $\boldsymbol{a}_m$ for the regression tree, we determine them so that the function defined by the regression tree approximates a gradient with respect to the current function $F_{m-1}(\boldsymbol{x})$ in the sense of least-square error as follows:

$$\boldsymbol{a}_m = \underset{\boldsymbol{a}}{\operatorname{argmin}} \sum_{i=1}^{N} \left(\tilde{y}_{im} - h\left(\boldsymbol{x}_i, \boldsymbol{a}\right)\right)^2, \tag{3}$$

where $\tilde{y}_{im}$ is the gradient and is given by

$$\tilde{y}_{im} = - \left[\frac{\partial \Psi(y_i, F(\boldsymbol{x}_i))}{\partial F(\boldsymbol{x}_i)} \right]_{F(\boldsymbol{x}) = F_{m-1}(\boldsymbol{x})}. \tag{4}$$

When the m-th regression tree using the $\boldsymbol{a}_m$ has L_m leaf nodes, the regression tree is given by

$$h(\boldsymbol{x}, \{R_{lm}\}_{l=1}^{L_m}) = \sum_{l=1}^{L_m} \bar{y}_{lm} 1(\boldsymbol{x} \in R_{lm}), \tag{5}$$

where R_{lm} is a disjoint region that the l-th leaf node of the m-th regression tree defines. $1(\cdot)$ is a Boolean function that outputs 1 in case the argument of the function is true. $\bar{y}_{lm}$ is a constant for the R_{lm}-th region, defined as the mean of training data that belongs to the l-th leaf node of the m-th regression tree. Since the output $\bar{y}_{lm}$ of the regression tree is constant, the weight β_m can be straightforwardly estimated using a line search on the loss function. Then, a new additive function $F_m(\boldsymbol{x})$ is updated as follows:

$$F_m(\boldsymbol{x}) = F_{m-1}(\boldsymbol{x}) + \nu \sum_{l=1}^{L_m} \gamma_{lm} 1(\boldsymbol{x} \in R_{lm}), \tag{6}$$

where $\gamma_{lm} = \beta_m \bar{y}_{lm}$ and $0 < \nu < 1$ is a shrinkage parameter to improve the generalization capability. In this study, we utilize the following least-square loss function $\Psi(y, F) = (y - F)^2/2$. This loss function leads to $\tilde{y}_{im} = y_i - F_{m-1}(\boldsymbol{x}_i)$ and $F_0(\boldsymbol{x}) = \bar{y}$. Here $\bar{y}$ is the mean of all the training data.

3. Experiments

3.1. Experimental Conditions

We conducted objective evaluations of the prediction error of Japanese, Mandarin, and English phone duration. The Japanese speech database consisted of 503 phonetically balanced sentences. These sentences were uttered by a female speaker and a male speaker in a normal reading style. The Mandarin speech database consisted of 1,680 sentences from the travel domain. These sentences were uttered by a female speaker in a normal reading style. These speakers are professional narrators. Detailed description and analysis of these database are given in [10],[11]. The English speech database used was *CSTR US KED Timit*[2], consisting of 453 phonetically balanced sentences. These sentences were uttered by an American male speaker in a normal reading style.

In the following experiments, we used the manually labeled phone duration and the following explanatory variables of the utterances of each speaker. The 47 Japanese explanatory variables consisted of 5 phonetic features, 3 mora-level features, 12 morpheme features, 12 accentual features, 12 breath-group-level features, and 3 utterance-level features. The 47 Mandarin explanatory variables consisted of 5 phonetic features, 10 tone relevant features, 12 morpheme features, 16 breath-group-level features, and 4 utterance-level features. The 53 English explanatory variables consisted of 5 phonetic features, 2 segment-level features, 22 syllable-level features, 12 word-level features, 9 phrase-level features, and 3 utterance-level features. Here the number of Japanese phonemes including vowel and consonant was 40. The number of Mandarin units including *initial*, which is an initial consonant of a syllable, and *final*, which is the part after the *initial* of the syllable,

[2] `http://festvox.org/dbs/dbs_kdt.html`

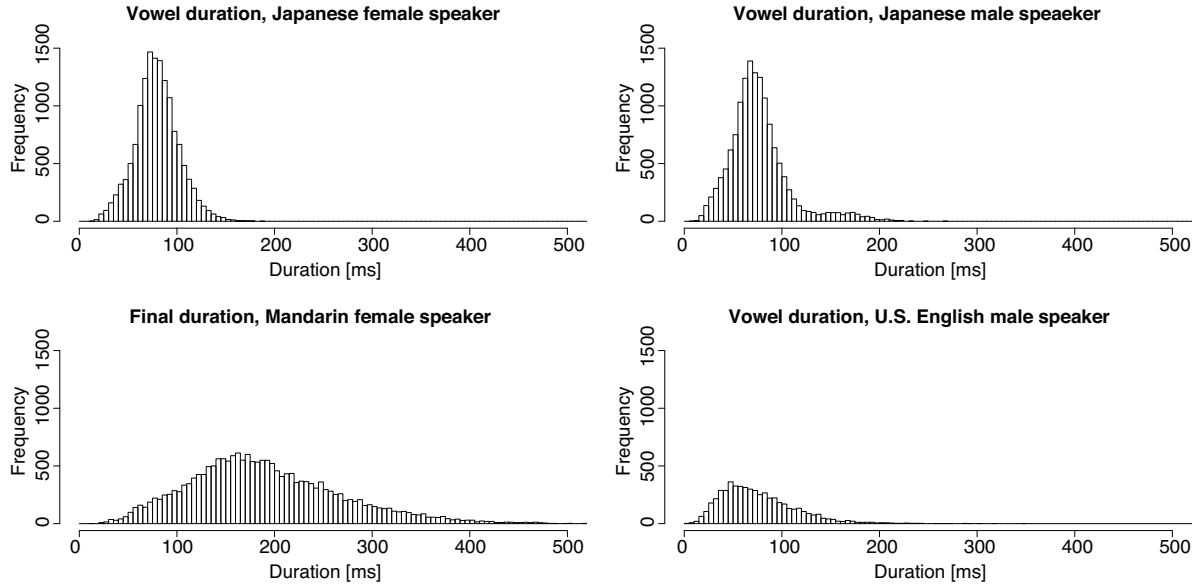

Figure 1. Histogram of vowel and *final* duration.

was 58[3]. The number of English phonemes including vowel and consonant was 44. Note that the accentual features in Japanese and tone-relevant features in Mandarin were also manually labeled based on the speech data. Figures 1 and 2 show the histograms of phone duration for these speakers. From these figures, we can clearly see that the *final* has a relatively longer duration than those of the Japanese or English vowels. This would be due to the fact that Mandarin *finals* consists of an optional final-head (medial), which is the diphthong glide before the center vowel, a final-center (nucleus), and an optional final-tail (coda), which is the part after the center vowel such as /-n/. Those optional parts in the *finals* can make its duration longer.

In addition to performance comparison of the GTB algorithm with the conventional duration modeling methods, there are a lot of interesting factors that might depend on the performance of the duration modeling. However, since manually labeling duration and several explanatory variables of the utterances requires huge costs, it is costly impractical to investigate all possible combinations of the factors. Thus, the following strategy was adopted for efficiency: Several factors and concerns such as gender or domain dependency in the GTB algorithm were first evaluated and analyzed in the Japanese database. Then, language dependency was evaluated via the comparison of GTB and other methods in all the languages.

We objectively evaluated all the sentences included in these speech databases using a 5-fold cross-validation method. For the objective evaluations of duration modeling, we utilized the following two measures: pseudo R-squared (R^2)[4]and root mean squared error

[3]Traditional descriptions of the Mandarin speech utilizes the *initials* and *finals* rather than individual phonemes.

[4]In nonlinear regression methods, strictly speaking, the original R-squared, that is, the square of the Pearson product-moment correlation coefficient, is not equivalent to (1 - the relative squared error).

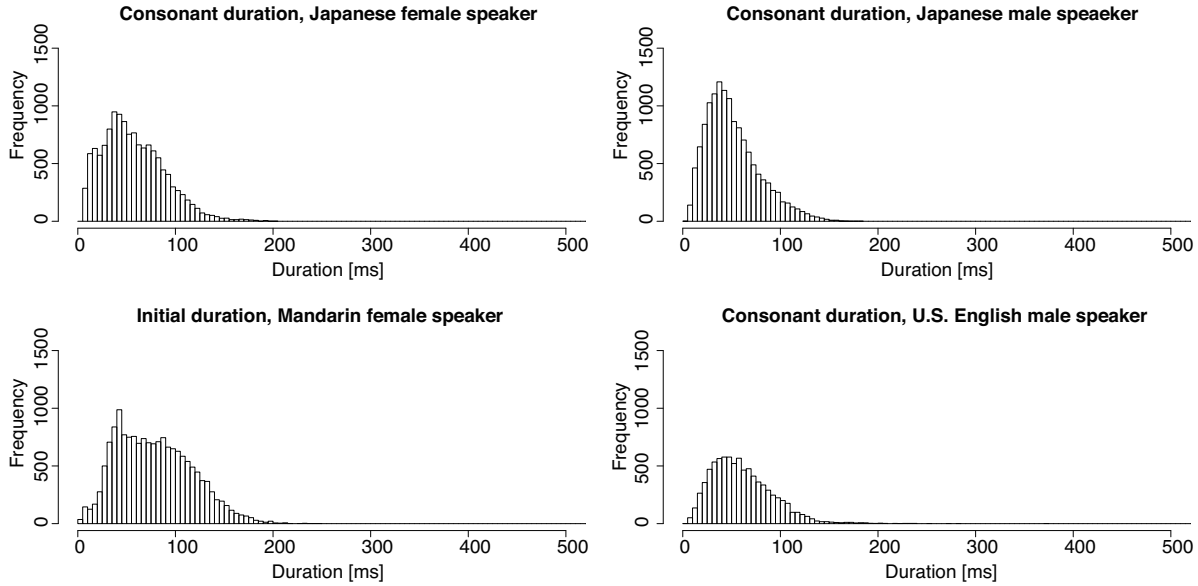

Figure 2. Histogram of consonant and *initial* duration.

(RMSE), between the predicted duration and that of the real utterance. The R^2 and RMSE are defined by

$$R^2 \equiv 1 - \frac{\sum_{i=1}^{T}(F(\boldsymbol{x}_i) - y_i)^2}{\sum_{i=1}^{T}(y_i - \bar{y})^2}, \tag{7}$$

$$\mathrm{RMSE} \equiv \sqrt{\frac{\sum_{i=1}^{T}(F(\boldsymbol{x}_i) - y_i)^2}{T}}, \tag{8}$$

where T is the number of test samples, and $\bar{y} = \sum_{i=1}^{T} y_i/T$ is the mean of the test samples.

We made use of a free data mining software called *WEKA*[5] [9] in these evaluations. For the GTB, we utilized binary regression trees. Note that we conducted pruning to the respective regression trees. We then set the number of regression trees M to 10 and the shrinkage parameter ν in the GTB to 0.5 based on several preliminary experimental results, respectively.

3.2. Evaluation of Phone Duration Modeling Using Gradient Tree Boosting

First, we evaluated the prediction accuracy on Japanese phone duration modeling. In each training of the cross-validation, we divided the set into two groups, that is, a vowel group and a consonant group, to assess the effect for each group. The regression trees were independently built for each group of each speaker. We also evaluated a conventional approach using regression trees as a baseline model.

Tables 1 and 2 show the results for the Japanese female and male speakers, respectively. In these tables, (a) shows the results of vowel duration, and (b) shows those of consonant duration. From the tables, we can firstly see that the conventional method using the

[5] `http://www.cs.waikato.ac.nz/ml/weka/`

Table 1
Objective evaluation results for a Japanese female speaker.

(a) Vowel

Model	R^2	RMSE (ms)
Regression Tree	0.55	14.94
GTB	0.61	13.87

(b) Consonant

Model	R^2	RMSE (ms)
Regression Tree	0.80	13.85
GTB	0.85	12.08

Table 2
Objective evaluation results for a Japanese male speaker.

(a) Vowel

Model	R^2	RMSE (ms)
Regression Tree	0.69	17.16
GTB	0.73	16.12

(b) Consonant

Model	R^2	RMSE (ms)
Regression Tree	0.73	14.24
GTB	0.78	12.77

regression trees has good R^2 values in both vowel and consonant duration as reported in the previous studies. Furthermore, we can see that the GTB algorithm improves the R^2 values and RMSE for both vowel and consonant duration, compared to the regression tree, that is, the baseline model of this method. Especially in consonant duration, we obtained substantial improvements. For example, compared with the regression tree, it reduced the RMSE of consonant duration for the female speaker from 13.85 [ms] to 12.08 [ms]. We attribute this to the ability of the GTB algorithm to capture the complex nonlinear functions with a high degree of accuracy.

3.3. Discussion

In this subsection, we explore the reasons the GTB improved the objective measures in detail. We inferred from several preliminary experiments that the following two advantages of the GTB are pivotal elements: 1) When the GTB constructs a new regression tree from the residuals of the current prediction function, this algorithm constructs it regardless of the structures of previous regression trees. As a result, in the space defined by the new tree, several errors that were calculated from different leaf nodes of the previous tree are merged into a single node. In other words, a leaf node of the new regression tree for the residuals effectively affects parts of several leaf nodes of the previous regression tree. On the other hand, the node splits of the regression tree only divide the disjoint regions into two smaller or more disjoint regions. 2) The shrinkage parameter ν in Eq. (6) has an ability to improve the generalization capability.

To confirm this, we compared the GTB with regression trees having almost the same number of leaf nodes. When the binary regression tree has $(L-1)$ nodes and L leaf nodes, the tree requires memory space for $L-1$ questions and indexes from the nodes to their child nodes, and requires model parameters of L constant values for the leaf nodes. On the other hand, when the GTB has M regression trees and each tree has L_m leaf nodes, it totally has $\sum_{m-1}^{M} L_m - M$ nodes and $\sum_{m-1}^{M} L_m$ leaf nodes. Since the weights β_m for the GTB are saved as γ_{lm} in the leaf nodes of the trees as shown in Eq. (6), the GTB only requires memory space for $\sum_{m-1}^{M} L_m - M$ questions and indexes to child nodes, and

Table 3
Comparison of numbers of leaf nodes of the regression tree and gradient tree boosting.

Model	Number of leaf nodes of the m-th regression tree											R^2	RMSE (ms)
	1	2	3	4	5	6	7	8	9	10	Sum		
Regression Tree	67										67	0.50	15.59
Regression Tree	166										166	0.52	15.21
Regression Tree	321										321	0.50	15.48
Regression Tree	498										498	0.49	15.66
GTB $\nu = 1.0$	67	40	32	24	1						164	0.55	14.71
GTB $\nu = 0.5$	67	71	59	33	34	22	17	20	1		324	0.58	14.18
GTB $\nu = 1.0$	166	67	27	29	4	12	14	1			320	0.55	14.71
GTB $\nu = 0.5$	166	93	59	58	32	32	21	17	18	1	497	0.59	14.03

model parameters of $\sum_{m-1}^{M} L_m$ constant values. Therefore, when the GTB has the same number of leaf nodes as that of the regression trees, i.e., $\sum_{m-1}^{M} L_m = L$, we can consider that both techniques have the same number of model parameters, and that the memory space for the GTB is slightly smaller than that for the regression tree.

We also evaluated objective measures using the GTB between before and after adjustment of the shrinkage parameter ν. We utilized a set of vowel duration of 400 sentences uttered by a Japanese female speaker as training data, and the other 103 sentences uttered by the same speaker as test data.

Table 3 shows the results of vowel duration. Furthermore, it shows the number of leaf nodes of each regression tree in the GTB and its sum. It also shows the results of only the first regression tree of the GTB and those of the regression trees having almost the same numbers of leaf nodes as those of the GTB. When we compare the several results of the regression trees, we can see that the regression tree having 166 nodes had the best objective measures, and the results of the trees having more than 166 nodes became worse because of overfitting. Then, from the comparison of the GTB using a total of 164 leaf nodes with the regression trees having 67 and 166 leaf nodes, it can been seen that the GTB using a total of 164 leaf nodes improved the objective measures to be better than the regression tree using 166 leaf nodes. For example, it reduced the RMSE from 15.59 [ms] to 14.71 [ms], whereas the regression tree only reduced it from 15.59 [ms] to 15.21 [ms]. Additionally, from the comparison of the GTB using a total of 320 leaf nodes with the regression trees having 166 and 321 leaf nodes, we see that the GTB similarly improved the objective measures, whereas the regression trees made them worse. Note that the shrinkage parameter ν in this GTB was set to 1, and the regression trees for the GTB were constructed directly from the residuals. Therefore, the only difference between these regression trees and the GTB is the splitting method for the nodes described above, and we can clearly see the effect from these results.

Moreover, from the comparison of the results using the experimentally adjusted shrinkage parameter ($\nu = 0.5$) with non-adjusted results ($\nu = 1$), we can see that it improved the objective measures further. The role of the shrinkage parameter is similar to a so-

Table 4
Comparison results with other duration modeling techniques for the Japanese female speaker.

(a) Vowel

Model	R^2	RMSE (ms)
Linear Regression	0.58	14.32
Regression Tree	0.55	14.94
Model Tree	0.58	14.28
Bagging Tree	0.57	14.51
GTB	0.61	13.87

(b) Consonant

Model	R^2	RMSE (ms)
Linear Regression	0.76	15.37
Regression Tree	0.80	13.85
Model Tree	0.83	12.94
Bagging Tree	0.82	13.47
GTB	0.85	12.08

Table 5
Comparison results with other duration modeling techniques for the Japanese male speaker.

(a) Vowel

Model	R^2	RMSE (ms)
Linear Regression	0.71	16.57
Regression Tree	0.69	17.16
Model Tree	0.72	16.48
Bagging Tree	0.71	16.71
GTB	0.73	16.12

(b) Consonant

Model	R^2	RMSE (ms)
Linear Regression	0.70	14.88
Regression Tree	0.73	14.24
Model Tree	0.76	13.53
Bagging Tree	0.74	13.88
GTB	0.78	12.77

called "learning rate" or "step length" parameter in steepest/gradient descent methods [12]. Then, from the comparison of the GTB ($\nu = 0.5$, total leaf nodes: 324) with the GTB ($\nu = 1$, total leaf nodes: 320), we can guess their combination effects. We have thus identified two reasons for which the GTB improved the objective measures to be better than the regression trees.

3.4. Comparison with Other Techniques

Next, we compared the GTB technique with several conventional techniques for duration modeling. Here we first selected the following three techniques related to the regression trees: linear regression [1], *model tree* [13], which is an integrated technique of the linear regression and the regression tree, and the *bagging* algorithm [14][15], which is a different meta algorithm of the regression trees.

In the linear regression, explanatory variables having multicollinearity were removed in advance. Then backward stepwise selection using the AIC criterion was used for eliminating unnecessary explanatory variables from the estimation of the regression coefficients. In the model tree, each leaf node utilized a linear regression function instead of a constant value. This approach is technically similar to [16]. Its training condictions used were the same as those of the regression trees and the above linear regression. In the duration modeling using the bagging algorithm [15], several pseudo sets of training data were created by using random sampling with replacement. Then, a regression tree was constructed

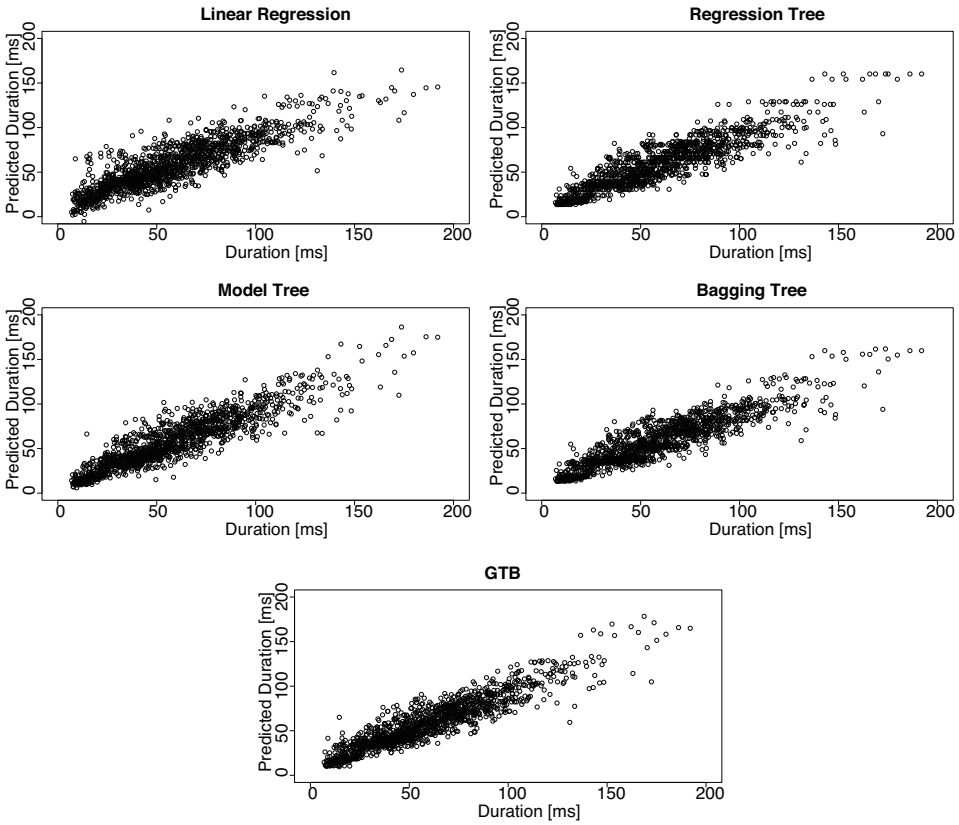

Figure 3. A scatter plot of the consonant duration for the Japanese female speaker.

for each set of training data. Finally, the average of the regression values of the several regression trees was utilized for determining the phone duration. This technique aims to obtain a more reliable predictive value than a single regression tree. For this bagging algorithm, the same number of trees as the GTB algorithm was used. Other experimental conditions are the same as in Section 3.2.

Tables 4 and 5 show the comparison results for the Japanese female and male speakers, respectively. In these tables, (a) shows the results of vowel duration, and (b) shows those of consonant duration. Comparing the results of the linear regression in these tables and those of the regression tree, we can see that the linear regression has better objective measures in vowel duration, whereas the regression tree has better objective measures in consonant duration. Thus, it can be seen that for both vowel and consonant duration, the model tree is slightly better than either the linear regression or regression tree. This is because the model tree can utilize the advantages of both the regression tree and linear regression. Furthermore, we can also see that the bagging algorithm has some effects on the objective measures compared to the regression tree. However, they were slight improvements. In addition to this, we can finally see that the GTB algorithm is better than these techniques in all the objective measures of both vowel and consonant duration.

Figure 3 shows the scatter plot of the real and predicted consonant duration for the Japanese female speaker. We can see that the plotted data of each technique diagonally distributes in this figure. Then, we can confirm that the conventional method using even

Table 6
Ratio of samples having allowable margin of errors and outliers of each technique. The consonant duration for the Japanese female speaker is used for evaluation. Standardized residuals ϵ of consonant duration are compared at several normal distribution percent points.

Model	Ratio of Acceptable Samples (%)			Ratio of Outliers(%)		
	$\|\epsilon\| < 0.03$	$\|\epsilon\|$ <0.13	$\|\epsilon\|$ <0.32	$\|\epsilon\|$ >1.96	$\|\epsilon\|$ >2.58	$\|\epsilon\|$ >3.29
$\mathcal{N}(0,1)$	2.0	10.0	25.0	5.0	1.0	0.1
Linear Regression	3.0	11.8	30.5	5.4	2.3	0.9
Regression Tree	2.8	13.5	29.5	5.3	1.5	0.7
Model Tree	2.8	13.3	32.7	5.6	2.0	0.8
Bagging Tree	2.8	13.4	30.2	5.3	1.5	0.7
GTB	3.5	13.3	31.8	5.4	1.6	0.7

Table 7
Comparison results with multilayer perceptrons for the Japanese male speaker.

(a) Vowel

Model	R^2	RMSE (ms)
MLP	0.54	22.65
GTB	0.73	16.12

(b) Consonant

Model	R^2	RMSE (ms)
MLP	0.55	20.22
GTB	0.78	12.77

the linear regression or regression trees can predict good duration to a certain degree. Moreover, we can also see that the GTB algorithm reduces the prediction errors and fits the data much better, since the distribution at the diagonal line has narrower breadth and higher density than other techniques.

Table 6 shows ratio of the samples having allowable margin of errors and outliers of each technique. The same consonant duration for the Japanese female speaker is used for the evaluation. Standardized residuals of the consonant duration are compared at several normal distribution percent points. For the samples having allowable margin of errors, the number of samples having standardized residuals under 0.03, 0,13 and 0.32 are counted. They correspond to 2%, 10%, and 25% points of the normal distribution. For the outliers, the number of samples having standardized residuals over 1.96, 2.58 and 3.29 are counted. They correspond to 5%, 1%, and 0.1% points of the normal distribution. From this table, we can confirm that the GTB algorithm reduces the prediction errors and generally increases the ratio of the samples having allowable margin of errors compared to the regression tree, and that linear regression have more outliers than other techniques. Unfortunately, the GTB algorithm does not seem to have an ability to reduce the outliers compared to the regression tree.

For reference, we compared the GTB algorithm with a neural-network-based method. As a feedforward neural network, we selected multilayer perceptrons (MLP). The MLP consisted of an input layer, a hidden layer having several units with sigmoid output functions, and an output layer which outputs weighted sum of the sigmoid functions. Its

Table 8
Objective evaluation results in different domains for a Japanese male speaker.

(a) Newspaper

Model	R^2	RMSE (ms)
Linear Regression	0.71	16.02
Regression Tree	0.73	15.47
Model Tree	0.75	14.83
Bagging Tree	0.74	15.11
GTB	0.77	14.38

(b) Travel conversation

Model	R^2	RMSE (ms)
Linear Regression	0.74	16.31
Regression Tree	0.74	15.78
Model Tree	0.75	15.33
Bagging Tree	0.75	15.27
GTB	0.78	14.61

Table 9
Objective evaluation results of a Mandarin female speaker.

(a) *Final*

Model	R^2	RMSE (ms)
Linear Regression	0.56	51.46
Regression Tree	0.56	51.40
Model Tree	0.56	51.44
Bagging Tree	0.59	49.26
GTB	0.64	46.19

(b) *Initial*

Model	R^2	RMSE (ms)
Linear Regression	0.76	18.75
Regression Tree	0.77	18.48
Model Tree	0.77	18.17
Bagging Tree	0.78	17.96
GTB	0.79	17.29

training algorithm used was the back-propagation. The number of epochs was 500. We first adjusted the number of the units in the hidden layer and then adjusted learning rate manually.

Table 7 shows the results of MLP and GTB. From the table, we can see that the MLP-based method has significantly worse results than those of the GTB algorithm. Considering even the simple linear regression or regression trees work well to a certain degree, this would be due to "Occam's razor," i.e., *entities should not be multiplied beyond necessity.*

3.5. Evaluation of Domain Dependency

Next, we investigated the prediction accuracy of the GTB algorithm and other related modeling techniques in different domains. The selected domains were newspapers and travel conversation. Texts used for the domains were selected from the Mainichi newspaper corpus[6] and the ATR Basic Travel Expression Corpus (BTEC) [17]. We utilized a speech database which contains 832 sentences for the newspapers domain and 1398 sentences for the travel conversation domain. These sentences were uttered by the Japanese male speaker in reading style, and phone duration and 47 explanatory variables of the utterances were manually labeled as heretofore. We utilized the same 5-fold cross validation method and evaluated all the sentences included in the database. In this experiment, we constructed the regression trees for each domain regardless of vowel and consonant.

[6] `http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html`

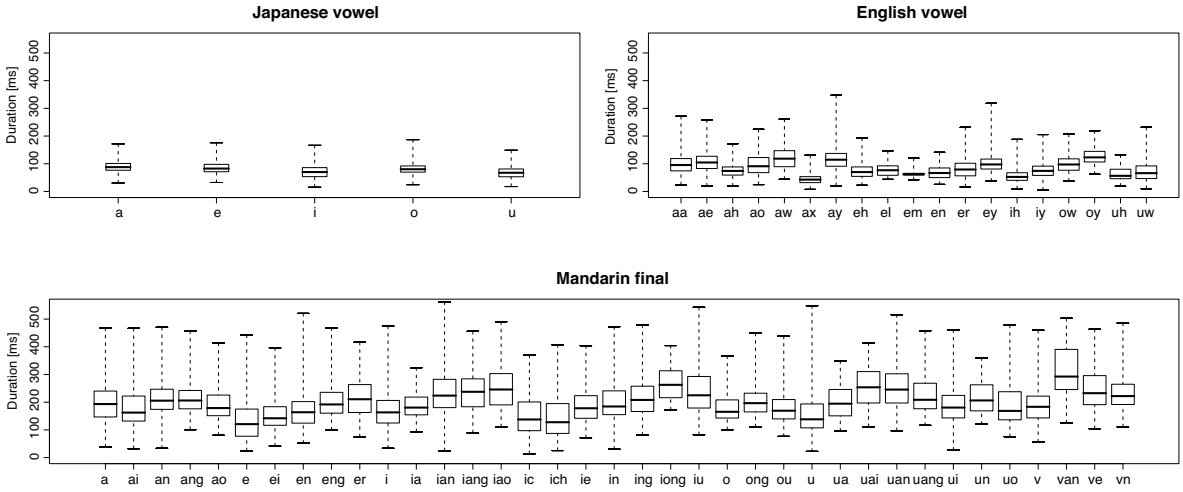

Figure 4. Comparison of vowel and *finals* duration of Japanese, Mandarin, and English.

Table 10. Ratio of samples having allowable margin of errors and outliers of each technique. The *final* duration for the Mandarin female speaker is used for evaluation. Standardized residuals ϵ of *final* duration are compared at several normal distribution percent points.

Model	Ratio of Acceptable Samples (%)			Ratio of Outliers(%)		
	$\lvert\epsilon\rvert < 0.03$	$\lvert\epsilon\rvert$ <0.13	$\lvert\epsilon\rvert$ <0.32	$\lvert\epsilon\rvert$ >1.96	$\lvert\epsilon\rvert$ >2.58	$\lvert\epsilon\rvert$ >3.29
$\mathcal{N}(0,1)$	2.0	10.0	25.0	5.0	1.0	0.1
Linear Regression	2.5	11.0	26.9	5.4	1.8	0.5
Regression Tree	3.1	12.1	28.4	6.0	1.9	0.4
Model Tree	2.4	10.0	25.6	5.7	1.7	0.4
Bagging Tree	2.8	12.8	28.7	5.6	1.8	0.5
GTB	3.2	12.2	29.7	6.0	1.9	0.5

Other conditions are the same as subsection 3.4.

Table 8 shows the results in these domains. In the table, (a) shows the result in the newspapers domain and (b) shows that in the travel conversation domain. These results indicate the same tendency of Table 5 and confirm again that the GTB algorithm outperforms all the other methods as well as previous experiments. It is also important to remember that the GTB algorithm is a robust meta algorithm which works well even in these different domains or conditions.

3.6. Evaluation of Language Dependency

Next, we evaluated the prediction accuracy on Mandarin phone duration modeling. In each training of the cross-validation, we divided the set into two groups, that is, an *initial* group and a *final* group. Other conditions are the same as in Section 3.2.

Table 9 shows the results of the Mandarin phone duration modeling. In the table, (a)

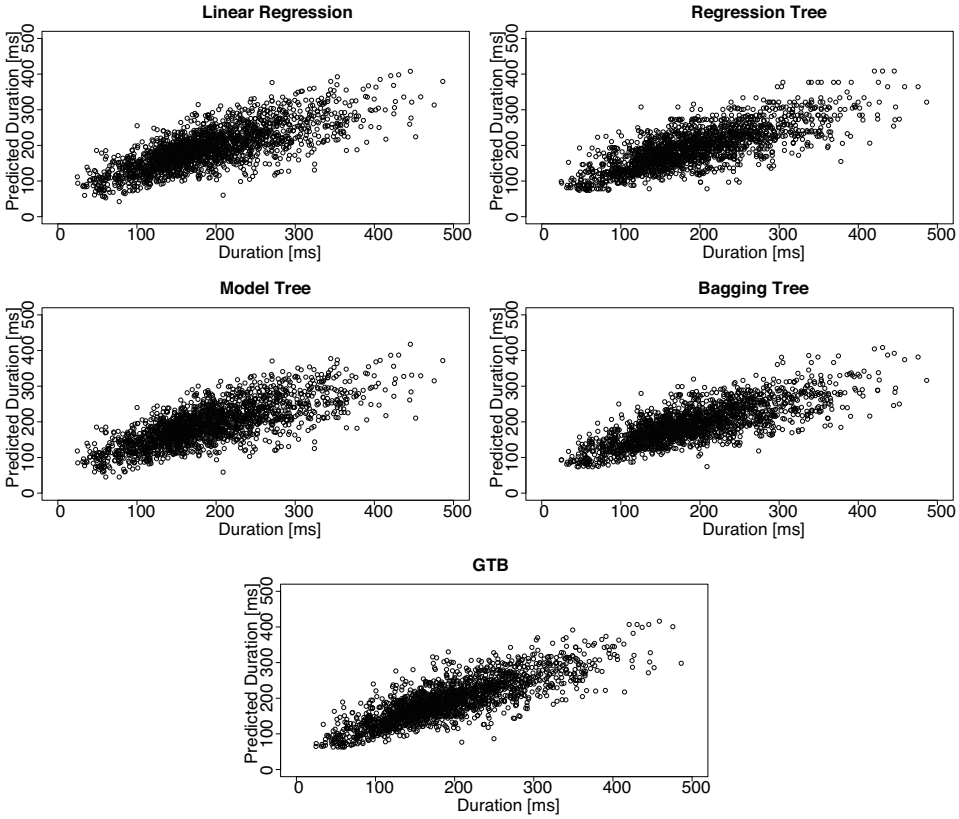

Figure 5. Scatter plot of the *final* duration for the Mandarin female speaker.

shows the results of *final* duration, and (b) shows those of *initial* duration. It can be seen that the GTB algorithm outperforms the conventional method using regression trees in both the *final* and *initial* parts. Contrary to the results for Japanese duration, we obtained substantial improvements in the *final* part. For example, it reduced the RMSE in the *final* part from 51.40 [ms] to 46.19 [ms]. In addition to this, we see that there seem to be differences between the characteristics of Mandarin phone duration in the *final* part and that of Japanese phone duration in the vowel part. For example, the RMSE is higher than that observed for Japanese phone duration. However, this is because the duration of each *final* in the Mandarin speech originally had larger variance than in Japanese or English speech as shown in Fig. 4. The figure shows the distributions of vowel and *final* duration used. The box-plots in the figure represents the shortest duration, lower quartile, median, upper quartile, and longest duration. From this figure, we can see that the *finals* originally have more than twice as long duration as the Japanese vowels. The same natural tendency of the much larger variation of the *final* duration can be seen in [18]. In fact, when we compare the accuracy of Mandarin and Japanese phone duration modeling in the R^2 where mean squared error is normalized by the variance, the accuracy of the Mandarin duration modeling is comparable to that of Japanese duration modeling.

Figure 5 and table 10 show the scatter plot of the real and predicted *final* duration, and ratio of the samples having allowable margin of errors and outliers of each technique in the *final* duration. From these results, we can confirm that the GTB algorithm reduces

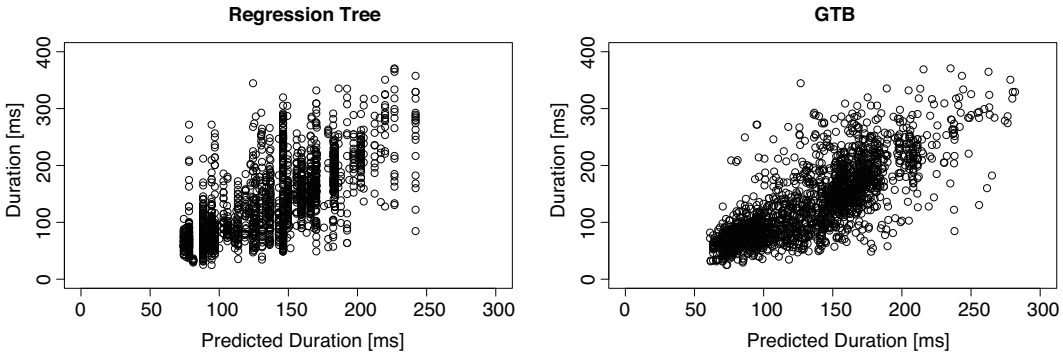

Figure 6. Scatter plot of the *final* "e" for the Mandarin female speaker.

Table 11
Objective evaluation results of an U.S. English male speaker.

(a) Vowel

Model	R^2	RMSE (ms)
Linear Regression	0.58	25.16
Regression Tree	0.54	26.41
Model Tree	0.58	25.35
Bagging Tree	0.56	25.82
GTB	0.61	24.51

(b) Consonant

Model	R^2	RMSE (ms)
Linear Regression	0.53	21.43
Regression Tree	0.53	21.51
Model Tree	0.55	21.06
Bagging Tree	0.55	20.94
GTB	0.58	20.18

the prediction errors and generally increases the ratio of the samples having allowable margin of errors even in the Mandarin speech data as well as the Japanese speech data. The differences between the GTB algorithm and regression tree in the figure and table seem smaller than those in the Japanese vowel duration. However, there are still crucial differences between the GTB algorithm and regression tree. Figure 6 show the scatter plot of the real and predicted duration of the *final* "e" for the female speaker. We can clearly see that the distribution of the predicted duration by the regression tree is very *discrete* and limited, since the outputs of the regression tree are restricted to the constants which leaf nodes have. On the other hand, we can see that the GTB algorithm overcomes the drawback of the regression tree and the distribution of the predicted duration by the GTB algorithm becomes continuous and varied, although the GTB algorithm is simple weighted sum of the multiple regression trees.

Finally, we evaluated the prediction accuracy on English phone duration modeling. Experimental conditions are the same as in Section 3.2. Table 11 shows the results of the English phone duration modeling. In this table, (a) shows the results of vowel duration, and (b) shows those of consonant duration. As in the previous experiments using Japanese and Mandarin speech data, we can see that the GTB algorithm substantially improves the objective measures of both vowel and consonant duration, compared to the regression tree. From these results, we can conclude that language dependency of the GTB algorithm is low and this algorithm would have similar effects in other languages.

4. Concluding Remarks

In this study, we incorporated the GTB algorithm into phone duration modeling as an alternative to the regression tree approach, and objectively evaluated the prediction accuracy of Japanese, Mandarin, and English phone duration. The GTB algorithm is a meta algorithm of regression trees: it iteratively builds the regression tree from the residuals and outputs *weighting sum* of the regression trees. Since it utilizes multiple regression trees, it increases the number of parameters compared to a single regression tree. However, the algorithm can robustly improve the prediction accuracy of the regression tree with ease, whereas increasing the number of the number of leaf nodes for the single regression tree put the accuracy at risk. As the same time, it solves a crucial problem of discrete outputs of the regression trees. In our experiments, it reduced the RMSE of consonant duration for the Japanese female speaker from 13.85 [ms] to 12.08 [ms] and the *final* duration for the Mandarin female speaker from 51.40 [ms] to 46.19 [ms]. Considering the building algorithms or criterion of each regression tree in the GTB algorithm are the same as those of the original regression tree, the amazing improvement rates of the GTB algorithm would be more than one expected. Then, it was constantly better than several techniques related to the regression trees and nonlinear regression techniques for duration modeling. Moreover, our evaluation results have confirmed that the GTB algorithm can substantially improve the predictive accuracy of the phone duration regardless of languages, speakers, or domains which we used. Since the information required for the GTB algorithm is the same as that for the conventional regression trees, the algorithm would be especially beneficial in a situation where automatically constructing the duration models is preferable to individually analyzing and determining the structure of the duration models, as in multilingual text-to-speech synthesis. Our future work will focus on F_0 modeling using ensemble learning.

Acknowledgments

The research reported here was developed at the Spoken Language Communication Research Laboratories of Advanced Telecommunications Research Institute International, and supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled “A study of speech dialogue translation technology based on a large corpus.”

The authors would like to thank Dr. Toshio Hirai of Arcadia Inc., Associate Professor Minoru Tsuzaki of Kyoto City University of Arts, Dr. Satoshi Nakamura and Dr. Jinfu Ni of ATR Spoken Language Communication Research Laboratories, Dr. Nobuyuki Nishizawa of KDDI R&D Laboratories, Professor Keiichi Tokuda of Nagoya Institute of Technology, and Dr. Korin Richmond of University of Edinburgh for their valuable discussions with us.

REFERENCES

1. K. Takeda, Y. Sagisaka, H. Kuwabara, 1989, On sentence-level factors governing segmental duration in Japanese, Journal of the Acoustic Society of America, 86, 6, 2081–2087.

2. M.D. Riley, 1992, Tree-based modelling of segmental duration, Talking Machines: Theories, Models, Designs, 265–273.
3. W. N. Campbell, 1990, Analog I/O nets for syllable timing, Speech Communication, 9, pp.57–61.
4. M. Riedi, 1995, A neural-network-based model of segmental duration for speech synthesis, In Proc. EUROSPEECH-95, pp.599–602.
5. J. van Santen, and J. Olive, 1990, The analysis of contextual effects on segmental duration, Computer Speech and Language, 4, 359-390.
6. J.H. Friedman, 2001, Greedy function approximation: A gradient boosting machine, Annals of Statistics, 29, 5, 1189–1232.
7. J.H. Friedman, 2002, Stochastic gradient boosting, Computational Statistics & Data Analysis, vol.38, no.4, pp.367–378.
8. T. Hastie, R. Tibshirani, and J.H. Friedman, 2001, The elements of statistical learning: data mining, inference, and prediction, Springer Series in Statistics, Springer.
9. I.H. Witten, E. Frank, 2005, Data Mining: Practical Machine Learning Tools and Techniques, Second Edition, Morgan Kaufmann.
10. H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, 2004, XIMERA: a new TTS from ATR based on corpus-based technologies, In Proc. 5th ISCA Speech Synthesis Workshop, 179–184.
11. H. Kawai, T. Toda, J. Yamagishi, T. Hirai, J. Ni, N. Nishizawa, M. Tsuzaki, and T. Tokuda, 2006, XIMERA: a concatenative speech synthesis system with large scale corpora, IEICE Trans., 2688-2698, J89-D-II, 12 (in Japanese).
12. J.A. Snyman, 2005, Practical mathematical optimization: An introduction to basic optimization theory and classical and new gradient-based algorithms, Springer.
13. J.R. Quinlan, 1992, Learning with continuous classes, In Proc. AI'92, 343–348.
14. L. Breiman, 1996, Bagging predictors, Machine Learning, 24, 123–140.
15. S. Lee, and Y. Oh, 1999, CART-based modelling of Korean segmental duration, In Proc. Oriental COCOSDA '99, 109–112.
16. N. Iwahashi, and Y. Sagisaka, 2000, Statistical modeling of speech segment duration by constrained tree regression, IEICE Trans. Information and Systems, E83-D, 7, 1550–1559.
17. T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, 2002, Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world, In Proc. of LREC 2002, 147–152.
18. S.H. Chen, W.H. Lai, and Y.R. Wang, 2003 A new duration modeling approach for Mandarin speech, IEEE Trans. on Speech and Audio Processing, 11, 4, 308–320.