Skeletonising Chinese Fundamental Frequency Contours with A Functional Model and Its Evaluation

Jinfu Ni & Hisashi Kawai

ATR Spoken Language Translation Research Laboratories, Japan

{jinfu.ni; hisashi.kawai}@atr.co.jp

Abstract

This paper presents a method for skeletonising a fundamental frequency (F_0) contour with its underlying F_0 peaks and valleys, without losing the linguistic and para-linguistic information that it conveys. The F_0 peaks and valleys are mainly associated with underlying lexical tones, and can be easily converted into other features, such as the response time and amplitude of local F_0 rise/fall movements. Consequently, the exact shape of the F_0 contour can be then recovered by the use of a functional F_0 model, given the F_0 peaks and valleys. Experiments were conducted on 668 Chinese utterances (around 1.4 hours of speech) from two native speakers. The validity of the proposed method is consistently proved by a three-fold evaluation: error analyses, perceptual similarity between the re-synthesised tone and intonation and the original, and a listening test of the naturalness of synthetic speech with incorporation of the recovered F_0 contours into the unit selection process for synthesis.

1. Introduction

Perception tests and instrumental analyses of the past have yielded a consensus that the fundamental frequency (F_0) contour of an utterance can multiply manifest lexical tones, stress and intonation [1][2][3][4]. Skeletonising F_0 contours is thus desirable in prosodic analysis and its application to speech information processing. The first reason for this is that the F_0 peaks and valleys play a prominent role in anchoring the tone and intonation patterns. Pitch targets, basically comprising *high* and *low*, are commonly used to describe the intonation of accent languages, such as English and Japanese [5]. In Chinese, however, there exist four lexical tones, named Tones 1 to 4, and a neutral tone named Tone 0. If the range of a speaker's voice is divided into four equal intervals, marked by five points, 1 low, 2 half-low, 3 middle, 4 half-high, and 5 high, Tones 1 to 4 are represented by 55, 35, 214, 51, respectively [1]. Because both the actual intervals and the absolute pitch are relative to the individual voice and the mood at the moment of speaking, the pitch targets are usually measured as F_0 peaks and valleys. Reliable analysis and labeling of the prosody must be capable of dealing with the tone variability under various conditions.

The second reason is related to the necessity of combining a statistical method with knowledge-based techniques to synthesise natural tone and intonation, arising from the development of text-to-speech conversion systems. Because the pitch targets can capture the interaction of the tone, stress and intonation [1], skeletonising F_0 contours shall be a key step in approaching such an aim. In this paper, we propose an efficient data-driven method upon our previous work to shrink an F_0 contour into the F_0 peaks and valleys that makes use of a functional F_0 model [6] [7]. This model bridges the gap between linguistic and acoustic F_0 features, and creates constraints to reduce speaker-dependent effects, thus facilitating the data-driven learning and parameter estimation.

The remainder of the paper explains this method. Section 2 includes a description of the model and the algorithm for skeletonising F_0 contours. Experimental results are described in Section 3, and remarks and future work are given in Section 4.

2. Outline of the method

It is commonly assumed that the F_0 contour of an utterance is the physical implementation of a sequence of discrete speech events or pitch targets through which the linguistic and para-linguistic information is conveyed. Because the vocal cords are a physical system, the F_0 contour produced by vocal cord vibrations is predicable to a certain extent, given the pitch targets. To bridge the gap between the acoustic and the linguistic features, a model is helpful for analysing and skeletonising the F_0 contours.

2.1. A functional model of the F_0 contours

In this paper, we use a functional model [6] to represent the observed F_0 contours in a parametric form. An advantage of this model, compared to the Fujisaki model [8], is that it supports automatic analysis of the F_0 contours [7]. According to the model, the voice register (a frequency register of utterances) of a speaker is first transposed to a so-called RONDO scale (similar to a log-scale). The RONDO-F_0 contour is then expressed in concatenative mountain-shaped patterns lined up in series at the time axis. The F_0 contour $F_0(t)$ is given as follows:

$$\frac{\ln F_0(t) - \ln f_{0_b}}{\ln f_{0_t} - \ln f_{0_b}} = \frac{A(\Lambda(t)) - A(\lambda_b)}{A(\lambda_t) - A(\lambda_b)}, \text{for } t \geq 0, \quad (1)$$

where

$$A(\lambda) = \frac{1}{\sqrt{(1-(1-2\zeta^2)\lambda)^2 + 4\zeta^2(1-2\zeta^2)\lambda}}, \lambda \geq 1, \quad (2)$$

and

$$\Lambda(t) = \Lambda_{r_1}(t) + \sum_{i=1}^{n-1} \text{Min}(\Lambda_{f_i}(t), \Lambda_{r_{i+1}}(t)) + \Lambda_{f_n}(t)\,. \quad (3)$$

$\text{Min}(z_1, z_2)$ means the smaller one of both z_1 and z_2. Equations (1) and (2) jointly indicate the transposition of the voice register, while Eq. (3) expresses the RONDO-F_0 contour $\Lambda(t)$, where $\Lambda_{r_i}(t)$ and $\Lambda_{f_i}(t)$ indicate the rise and fall components of the ith mountain-shaped pattern, respectively. Particularly,

$$\Lambda_{r_i}(t) = \begin{cases} \lambda_{p_i} + \Delta\lambda_{r_i}(1 - D_{r_i}(t_{p_i} - t)), & \text{for } t \leq t_{p_i}, \\ 0, & \text{otherwise,} \end{cases} \quad (4)$$

$$\Lambda_{f_i}(t) = \begin{cases} \lambda_{p_i} + \Delta\lambda_{f_i}(1 - D_{f_i}(t - t_{p_i})), & \text{for t} \geq t_{p_i}, \\ 0, & \text{otherwise,} \end{cases} \quad (5)$$

$$\text{where } D_{x_i}(t) = (1 + \frac{4.8t}{\Delta t_{x_i}})\, e^{-\frac{4.8t}{\Delta t_{x_i}}}, \text{for } t \geq 0. \quad (6)$$

Parameters ζ, λ_t and λ_b can be commonly fixed at 0.237, 1 and 2, respectively [6]. There are then two speaker-dependent but utterance-independent parameters in the frequency domain,

$[f_{0_b}, f_{0_t}]$: top and bottom frequencies of the voice register,

and five utterance-dependent but speaker-independent parameters in the RONDO-time space,

n : number of mountain-shaped patterns,
Δt_{x_i} : response time for the *i*th rise/fall component,
$\Delta\lambda_{x_i}$: amplitude of the *i*th rise/fall component, $x \in \{r, f\}$
(t_{p_i}, λ_{p_i}) : peak of the *i*th mountain-shaped pattern, $i = 1, ..., n$.

Figure 1 shows the tone modeling with the mountain-shaped patterns and the association of the model parameters with the mountain-shaped pattern, where H, R, L and F indicate Tones 1 to 4. Because the close correlation of the model parameters with the peaks and valleys of a tone, it is reasonable to use the functional model while skeletonising an F_0 contour with its underlying F_0 peaks and valleys.

2.2. Outline of the algorithm

Let us take the example shown in Fig. 2 to describe the process of skeletonising an F_0 contour and demonstrate the performance of this method. Given the observed F_0 contour (the "+" sequence) shown in Fig.2 (a), it is first represented in a parametric form based on the functional model using the method in [7]. Consequently, the peak resulting from a part of the set of model parameters, and a model-approximated contour is available, given the estimated parameters. Figure 2 (b) shows the model-approximated contours (the solid lines). Then, a valley is searched for the rise/fall components around the peak in the RONDO-time space. Figure 2 (c) shows the peak (the solid circles) and the valleys (the empty circles), giving the skeleton of the F_0 contour. The F_0 peaks and valleys can be then converted into the model parameters to recover the F_0 contour; a copy is shown in Fig. 2 (d) (the solid lines).

2.2.1. Improved parameter estimation

A method has been proposed to extract the tone peak and gliding features from observed F_0 contours that makes use of the functional model [7]. According to this method, the tone peaks are first determined by adjusting several baseline tone patterns to fit the F_0 contour fragment of a syllable with the analysis-by-synthesis-based pattern matching technique. Tone gliding features are then re-estimated after the determination of tone peaks with the criterion of minimising the error between the model-approximated contours and the observed ones.

To reliably skeletonise the F_0 contours, a few constraints were newly incorporated into the algorithm for re-estimation of the model parameters relative to the tone gliding features, namely, Δt_{x_i} and $\Delta\lambda_{x_i}$. Let $(\hat{\lambda}_{v_i}, \hat{t}_{v_i})$ denote the observed F_0 valleys between the *i*th peak (λ_{p_i}, t_{p_i}) and the next, taking into account the voice frame probability to suppress the effect of potential F_0 extraction errors. The constraints for re-estimation of Δt_{x_i} and $\Delta\lambda_{x_i}$ are listed below.

$$\lambda_{p_i} + \Delta\lambda_{f_i} \leq \hat{\lambda}_{v_i} \times 1.1 \quad (7)$$
$$\lambda_{p_i} + \Delta\lambda_{f_i} = \lambda_{p_{i+1}} + \Delta\lambda_{r_{i+1}} \quad (8)$$
$$\Delta\lambda_{f_i} \geq 0.02 \quad (9)$$
$$\Delta\lambda_{r_{i+1}} \geq 0.02 \quad (10)$$
$$0.05 \leq \Delta t_{f_i} \leq (\hat{t}_{v_i} - t_{p_i}) \times 1.1 \quad (11)$$
$$0.05 \leq \Delta t_{r_{i+1}} \leq (t_{p_{i+1}} - \hat{t}_{v_i}) \times 1.1 \quad (12)$$

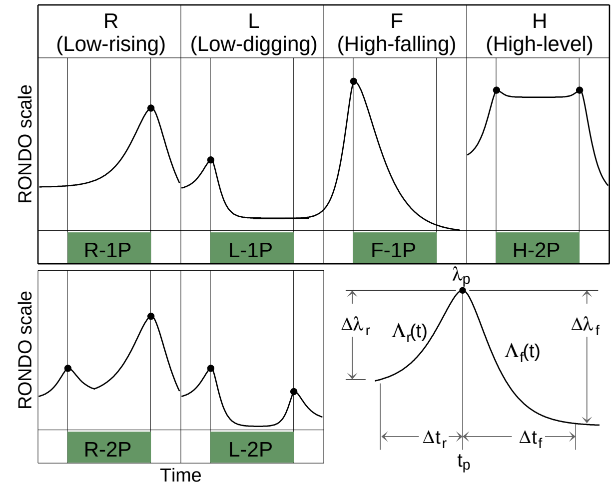

Figure 1: *Tone modeling with the mountain-shaped patterns. A mountain-shaped pattern with its control parameters is also superimposed on this figure. Solid circles indicate peaks.*

2.2.2. Valley search

Search of an F_0 valley $(t_{v_{x_i}}, \lambda_{v_{x_i}})$ for either of the *i*th rise and fall components is performed on the RONDO-contours around the *i*th peak. The candidate for a valley, for example, $(t_{v_{f_i}}, \lambda_{v_{f_i}})$ is first set at the valley $(\bar{t}_{v_i}, \bar{\lambda}_{v_i})$ of the RONDO-contours between the *i*th peak and the next. Then, the valley candidate is moved toward the *i*th peak through decreasing $t_{v_{f_i}}$ with a very short interval (e.g., 0.005 sec) along the RONDO-contour until

$$\lambda_{v_{f_i}} - \lambda_{p_i} \leq (\bar{\lambda}_{v_i} - \lambda_{p_i}) \times 0.95, \quad (13)$$

or $t_{v_{f_i}} = t_{p_i}$. In Eq. (13), the constant 0.95 is determined by considering the definition of Δt_x as the response time required for unit decay from 1 to 0.05, i.e.,

$$\mathrm{D}_x(\Delta t_x) = (1 + \alpha\Delta t_x)e^{-\alpha\Delta t_x} = 1 - 0.95. \quad (14)$$

The relationship between α and Δt_x can be expressed as

$$\alpha = \frac{4.8}{\Delta t_x}. \quad (15)$$

It is noted that if the difference between $t_{v_{f_i}}$ and $t_{v_{r_{i+1}}}$ is less than a threshold, as the two valleys lie between them.

2.2.3. Parameter conversion

Given the peaks and valleys, the other model parameters necessary for recovering an F_0 contour are calculated as follows:

$$\Delta t_{r_i} = \max(0.05, t_{p_i} - t_{v_{r_i}}), \quad (16)$$
$$\Delta\lambda_{r_i} = \max(0.02, (\lambda_{v_{r_i}} - \lambda_{p_i}) \times 1.05), \quad (17)$$
$$\Delta t_{f_i} = \max(0.05, t_{v_{f_i}} - t_{p_i}), \quad (18)$$
$$\Delta\lambda_{r_i} = \max(0.02, (\lambda_{v_{f_i}} - \lambda_{p_i}) \times 1.05). \quad (19)$$

3. Experimental evaluation

Two experiments were conducted on 668 Chinese utterances to test the effectiveness of this method. Experiment 1 rates the perceptual similarity between the recovered tone and intonation patterns and the original. Experiment 2 judges the naturalness of synthetic speech with the effects of the modeling (i.e., automatic parameter estimation) and the skeletonising on the prosodic properties by comparing them with the original. For reference, the average errors between the model-generated contours and the original were also calculated.

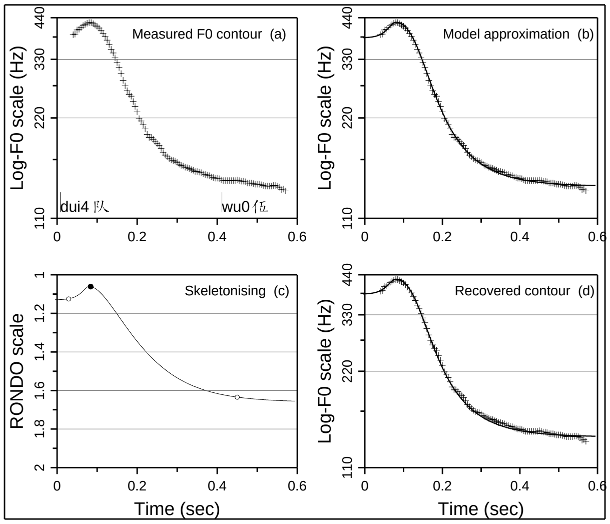

Figure 2: *Illustration of skeletonising an F_0 contour, where "+" sequence indicates the observed F_0 contour; the solid lines indicate the model-approximated contour; the full circle indicates the F_0 peak and the empty circles indicate the F_0 valleys.*

3.1. Experiment 1: Perceptual similarity between the re-synthesised tone and intonation patterns with the original

3.1.1. Speech material and methodology

The speech material used in this experiment includes 72 sentences, which are almost all adopted from [9]. These sentences are divided into six groups, each containing 12 base sentences further subdivided into three types. Each type includes four sentences of the same number of syllables and of the same grammatical structure characterized by the mapping of an identical tone onto the entire sentence. Type 1 comprises four syllables with subject-verb (SV) structures; Type 2 has five syllables with subject-verb-object (SVO) structures; and Type 3 combines Types 1 and 2, i.e., Type 2 was added to Type 1 as its sentential object, yielding nine syllables with SVO structures. These sentences are grouped into the following categories.

A: Types 1, 2, 3 in statements
B: Types 1, 2, 3 in lexically and grammatically unmarked yes -no questions
C: Types 1, 2, 3 in yes-no questions with interrogative particle *ma0* in sentence-final positions
D0: Type 2 in yes-no questions with *shi4-bu2-shi4* structures
D1: Type 2 in yes-no questions with *X-mei2-X* structures
D2: Type 2 in yes-no questions with *X-le0mei2-X* structures
E0: Type 2 in alternative questions with *X-hai2shi4-Y* structures
E1: Type 2 in questions with *shi4-X-hai2shi4-Y* structures
E2: Type 2 in questions with *hai2shi4-X-hai2shi4-Y* structures
F0: Type 2 in why (*wei4she2me0*) questions
F1: Type 2 in when (*she2me0shi2hou4*) questions
F2: Type 2 in what (*she2me0*) questions

We recorded the 72 sentences twice in a sound-proofed room with a female speaker without expressive emotion. The voice register of the speaker $[f_{0_b}, f_{0_t}]$ was fixed at [110 Hz, 500 Hz] for the parameter estimation. The 144 observed F_0 contours were first automatically analysed using the method, after which the F_0 peaks and valleys were manually checked with a visual inspection of the F_0 contours, taking into account the underlying tones. The number of F_0 peaks for each tone was basically determined according to the tone modeling shown in Figure 1. With model-generated F_0 contours, we re-synthesised the 144 utterances for perceptual experiments using a tool called STRAIGHT [10].

3.1.2. Results

Table 1 shows the statistical results of the tone-related samples measured from the speech material, where μ_c and σ_c indicate the mean and variance of these manually checked model parameters (*checked parameters*), respectively; μ_p and σ_p indicate those predicted by the F_0 peaks and valleys (*prediction parameters*). The columns μ_e and σ_e list the mean and variance of the errors between the checked and prediction parameters.

Table 1: *Statistical results of the model parameters.*

	Count	μ_c	σ_c	μ_p	σ_p	μ_e	σ_e
Δt_r	366	0.140	0.047	0.122	0.043	0.022	0.022
$\Delta\lambda_r$	366	0.224	0.147	0.215	0.143	0.013	0.035
Δt_f	382	0.139	0.047	0.134	0.055	0.019	0.027
$\Delta\lambda_f$	382	0.196	0.129	0.188	0.122	0.007	0.015

The average errors between the model-generated contours and the observed ones were 6.38 Hz (1.64 Hz per 100 Hz) for those with the prediction parameters and 5.94 Hz (1.52 Hz per 100 Hz) for those with the checked parameters, respectively.

To test the similarity between the model-generated tone and intonation patterns and the original, we performed a perceptual experiment with 288 stimulus pairs, including 144 re-synthesised utterances with the checked parameters and 144 utterances with the prediction parameters. The stimuli were presented to two native speakers through headphones in a silent room. After hearing each pair of stimuli, the listener rated the similarity of the tone and intonation between them with a three-point scale, 0 (very different), 1 (similar), 2 (no difference). The listeners were allowed to hear the same stimuli several times before making a judgment. The average scores for the checked and prediction parameters were 1.93 and 1.89, respectively, and no "very different" samples occurred. The experimental result indicated that the pitch targets, i.e., the F_0 peaks and valleys over time, suffice to capture the nature of the tone and intonation patterns.

3.2. Experiment 2: Application of the recovered F_0 contours to the unit selection for speech synthesis

Experiment 2 was conducted on 524 utterances from another speaker; the prosodic and spectral features extracted from these utterances were used as the targets to guide the unit selection for synthesis of the speech samples used in this experiment. The voice register of the speaker $[f_{0_b}, f_{0_t}]$ was fixed at [120 Hz, 420 Hz]. The speech samples were prepared in five steps.

Step 1: Extracting the F_0 contours from the 524 utterances and parameterising them based on the functional model.
Step 2: Skeletonising these F_0 contours with the F_0 peaks and valleys using the proposed method.
Step 3: Converting the F_0 peaks and valleys into the model parameters using Eqs. (16)-(19).
Step 4: Recovering the F_0 contours using these parameters.
Step 5: Corpus-based synthesis with the recovered F_0 contours.

An example of the skeleton of F_0 contours and F_0's recovered contours are shown in Figure 3. This example and all of the analyzed samples showed that the recovered F_0 contours closely

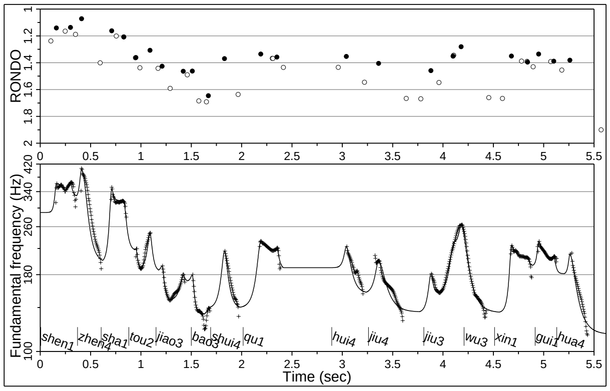

Figure 3: *Example of the skeleton of F_0 contours (top panel) and the recovered ones (the solid lines in the buttom panel). The solid and empty circles indicate the peaks and valleys, respectively. The "+" sequence indicates the observed F_0 cotours.*

Table 2: *Summary of the listening test results*.

Naturalness	Count	Skeleton (%)	Modelling (%)	Original (%)	Count	Same (%)
Improved	466	12.45	4.94	9.66	302	6.40
Degrade	466	11.16	4.72	3.00	302	2.65

matched the original. The average errors are 6.01 Hz (2.0 Hz per 100 Hz) for those with the prediction model parameters, and 3.63 Hz (1.21 Hz per 100 Hz) for those re-synthesised by the auto-estimated model parameters.

The speech corpus used for the speech synthesis consists of 20-hour speech data from one speaker, and the unit selection algorithm is an updated version of the suggestion [11]; no diphone unit was used here. There exist five sub-costs to rate the difference between a candidate and the target. This experiment only focused on the effect of the F_0 contours on the naturalness of synthetic speech, taking one of the three tokens in turn: the recovered F_0 contours (hereafter, *skeleton*), those with auto-estimated model parameters (*modelling*), and the observed F_0 contours (*original*). As a result, we obtained 524 stimuli; each consists of the three synthetic speech samples in a random order. The stimuli were presented to the two natives through headphones in a silent room. After hearing a set of stimuli, the listener was asked to rate the difference in naturalness among them and answer the two following questions.
Is there any difference in naturalness among the three samples? If different, which is the best or the worst?

The experimental results are summarised in Table 2. According to the output of the unit selection module, there are 466 sets of samples, in which there exists at least one different unit candidate among them. On the other hand, there are 302 pairs of samples with identical unit candidates for each pair. According to the result shown in Table 2, human perception may perceive identical samples with different perceptual impressions of naturalness: improved 6.4% and degraded 2.65%. Taking into account the perceptual errors, the results obtained from 466 sets of samples indicate that the recovered F_0 contours can capture the essential properties of the observed F_0 contours, as proved in Experiment 1.

4. Remarks and future work

This paper presents a method for skeletonising an F_0 contour with its underlying F_0 peaks and valleys that makes use of a functional F_0 model. Several analysis and perceptual experiments were conducted on the speech material designed for studying Chinese tone and intonation patterns and speech synthesis. Experimental results indicated that the pitch targets play a prominent role in anchoring the tone and intonation patterns; the exact F_0 contours can be predicted from the F_0 peaks and valleys using the functional model, without losing the primary linguistic and para-linguistic information that it conveys.

Future work will include applying this F_0 skeletonising method to speech information processing, such as investigation of a pitch-target-based method for analysing and synthesising the tone and intonation patterns to improve the naturalness of the synthetic speech.

Acknowledgement This research was supported in part by the Telecommunications Advancement Organization of Japan.

5. References

[1] Chao, Y. R., 1968. A Grammar of Spoken Chinese. Berkeley, University of California Press.

[2] Shen, J., 1994. Hànyŭyŭdiàohéyŭdiàolèixíng. *Zhōngguó yŭwén*, 3, 221-228.

[3] Kratochvil, P., 1998. Intonation in Beijing Chinese. *Intonation Systems, A Survey of Twenty Languages*, ed. by Hirst, D. and Cristo, A.D., Cambridge Uni. Press, 417-431.

[4] Wu, Z., 2000. From Traditional Chinese Phonology to Modern Speech Processing — Realization of Tone and Intonation in Standard Chinese —, *ICSLP2000*. Beijing.

[5] Beckman, M. E. and Pierrehumbert, J. B., 1986. Intonational Structure in Japanese and English, *Phonology Yearbook 3*, 255-309.

[6] Ni, J. and Hirose, K., 2000. Experimental Evaluation of a Functional Modeling of Fundamental Frequency Contours of Standard Chinese Sentences. *ISCSLP2000*. Beijing, 319-322.

[7] Ni, J. and Kawai, H., 2003. Tone Feature Extraction through Parametric Modeling and Analysis-by-Synthesis-based Pattern Matching. *ICASSP2003*. Vol. 1, 72-75.

[8] Fujisaki, H. and Hirose, K., 1984. Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese. *J. Acoust. Soc. Jpn (E)*, Vol.5, No.4, 233-242.

[9] Shen, X. S., 1990. The Prosody of Mandarin Chinese. University of California Publications.

[10] Kawahara, H., Ikuyo, M. K., Cheneigne, A., 1999. Restructuring Speech Representations Using a Pitch-Adaptive Time-Frequency Smoothing and an Instantaneous-Frequency-Based F0 Extraction: Possible Role of a Repetitive Structure in Sounds. *Speech Communication*, 27, 187-207.

[11] Toda, T., Kawai, H., Tsuzaki, M., and Shikano, K., 2002. Unit Selection Algorithm for Japanese Speech Synthesis Based on Both Phoneme Unit and Diphone Unit. *ICASSP2002*. 465-468.