STATISTICAL PHRASE/ACCENT COMMAND ESTIMATION ALGORITHM UTILIZING LINGUISTIC INFORMATION

Ryotaro Sato[1], *Kunio Kashino*[1,2]

[1] Graduate School of Information Science and Technology, The University of Tokyo
[2]NTT Communication Science Laboratories, NTT Corporation

ABSTRACT

The importance of extracting non-linguistic information has been highlighted in a growing variety of applications of speech signal processing. Among the audio features carrying such information, fundamental frequency (F_0) contours are considered primarily important. The Fujisaki model is a physical model that describes a F_0 contour with only a small number of parameters, namely, the timings and magnitudes of the phrase and accent commands, and a stochastic formulation and estimation algorithm have recently been proposed for it. However, the use of linguistic information has so far been limited, while it is known that accent commands are strongly related to linguistic information in many languages, and linguistic information could be obtained from the input audio signals by using speech recognition techniques. Against this background, this paper introduces a novel F_0 command parameter estimation method that incorporates linguistic information with the stochastic framework. Experiments using real speech data show that when linguistic information is appropriately utilized, the estimation accuracy of accent command parameters is improved by 43% under the proposed criteria.

Index Terms— voice fundamental frequency contour estimation, Fujisaki model, prosodic information processing, EM algorithm, speech recognition

1. INTRODUCTION

With the recent flourishing applications of speech signal processing technologies, the extraction and analysis of non-linguistic information have become more and more important [1]. Voice fundamental frequency (F_0) contours are physical entities reflecting such non-linguistic information as a speaker's identity, attitude, intention, and mood. An F_0 contour is typically represented by the sum of two components: a long-term varying component corresponding to a whole phrase, and a component that exhibits rapid changes according to the syllable units. The Fujisaki model [2] assumes that these two kinds of component are the output from a linear system. The input, called a "command", is an impulse-like signal for the long-varying component and a step-like signal for the rapidly changing one. In other words,

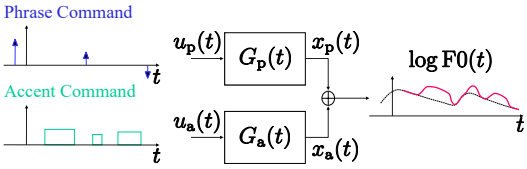

Fig. 1. Concept of Fujisaki model.

the Fujisaki model assumes that F_0 contours are reproduced by a series of *command parameters*, that is, the timings and values of the commands. Representing complex F_0 contours with such a highly compressed compact set of parameters is important not only for various applications such as mood recognition from speech and natural conversation systems, but also for scientific studies of prosody itself. Hence, it is worth establishing the technique to extract these parameters from speech signals with high accuracy.

Command parameter estimation for the Fujisaki model has been a long-standing research topic [3] because it is basically an ill-posed inverse problem. To meet this challenge, a generative model-based approach has recently been proposed [4]. From the viewpoint of this approach, integrating multiple pieces of information is expected to be useful to improve its estimation accuracy [5]. Since linguistic information is known to be tightly bound to accent information in many tone languages and pitch-accent languages, here we consider the use of linguistic information.

In the following sections, we first briefly review the framework. We then propose a new method for the Fujisaki model command parameter estimation by taking advantage of linguistic information contained in speech signals. In Section 4, we show that this model yields an improvement in estimation accuracy. A conclusion is presented in Section 5.

2. GENERATIVE MODEL BASED APPROACH TO F_0 PARAMETER ESTIMATION

2.1. Fujisaki model, describing the F_0 contour

The Fujisaki model [2] assumes that a speech F_0 pattern on a logarithmic frequency scale $y(t)$ is given as the sum of three components: $y(t) = x_{\mathrm{p}}(t) + x_{\mathrm{a}}(t) + \mu_{\mathrm{b}}$, where x_{p} and x_{a} represent a phrase and accent component, and μ_{b} is a constant offset value (Fig. 1). The phrase and accent components are

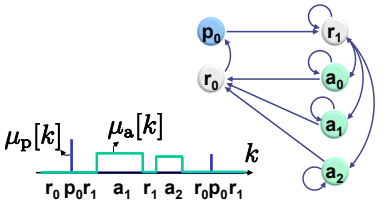

Fig. 2. HMM generating Fujisaki model command functions and graph of typical output parameters.

both assumed to be output from potentially different second-order critically damped filters, $G_{\mathrm{p}}(t)$ and $G_{\mathrm{a}}(t)$, which are excited with a pulse sequence $u_{\mathrm{p}}(t)$ (phrase commands) and a rectangular pulse sequence $u_{\mathrm{a}}(t)$ (accent commands), respectively.

This model can be viewed as a compression of F_0 contours to command functions sparsely valued in time. Therefore, they are expected to be useful when analyzing the relationship between prosody and individuality or emotion, for example.

The Fujisaki model is applicable to a wide range of languages. In a considerable number of tone languages and pitch-accent languages, linguistic information is known to be tightly bound to accent information and also to command functions. For example, in the case of Japanese, accents are considered to synchronously rise with linguistic accents.

2.2. Generative model of F_0 command parameters

Over the years, many researchers have proposed methods to automatically extract command parameters of the Fujisaki model [6]. Our method is based on the generative model approach introduced by Kameoka et al. [4].

This method introduces a path-restricted hidden Markov model (HMM) that outputs the Fujisaki model command parameters (Fig. 2). We denote the state transition sequence at each frame k by $\boldsymbol{s} = (s_1, \ldots, s_k, \ldots, s_K)$. The output consists of two-dimensional real values $\boldsymbol{o}[k] = (u_{\mathrm{p}}[k], u_{\mathrm{a}}[k])^\top$, representing the phrase/accent command amplitude at frame k. Output values are independently normally distributed, and the mean and variance of each component at each state are model parameters. The state p_0 means that only the phrase command is activated ($\mu_{\mathrm{p}}[k] := \boldsymbol{E}[u_{\mathrm{p}}[k]] = C^{(\mathrm{p})}[k], \mu_{\mathrm{a}}[k] := \boldsymbol{E}[u_{\mathrm{a}}[k]] = 0$). Similarly, at the states $\mathrm{a}_n (n = 0, \cdots, N-1)$, only the accent command is activated ($\mu_{\mathrm{p}}[k] = 0, \mu_{\mathrm{a}}[k] = C_n^{(\mathrm{a})}$), and both are deactivated ($\mu_{\mathrm{p}}[k] = \mu_{\mathrm{a}}[k] = 0$) at the states $\mathrm{r}_i (i = 0, 1)$. The path constraint restricts the sequence of μ_{p} to one consisting of isolated deltas and $\mu_{\mathrm{a}}[k]$ of rectangular pulses, which are important features of commands functions. In summary, given a sequence $\boldsymbol{s}$, this HMM emits $\boldsymbol{O} = (\boldsymbol{o}[1], \ldots, \boldsymbol{o}[K])$ according to the Gaussian distribution: $\boldsymbol{O} \sim \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{o}[k]; \boldsymbol{\rho}[k], \boldsymbol{\Upsilon})$, where $\boldsymbol{\rho}[k] = (\mu_{\mathrm{p}}[k], \mu_{\mathrm{a}}[k])^\top$ and $\boldsymbol{\Upsilon} = \mathrm{diag}(\sigma_{\mathrm{p}}^2, \sigma_{\mathrm{a}}^2)$. In addition, we introduce the conditional density $P(\boldsymbol{y}|\boldsymbol{O}, \mu_{\mathrm{b}})$ for log F_0 values $\boldsymbol{y} = (y[1], \ldots, y[K])$ using the Gaussian distribution.

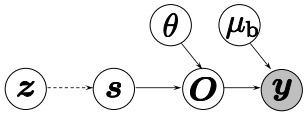

Fig. 3. Graphical representation of the generative model for F_0 command parameters. Here, the only visible parameter is $\boldsymbol{y}$. The dashed arrow from the variable $\boldsymbol{z}$ means that this variable exists only in the proposed method, not in the conventional ones.

2.3. Conventional estimation algorithm

By using the stochastic command parameter model, the joint probability density function can be written as $P(\boldsymbol{s}, \theta, \mu_{\mathrm{b}}, \boldsymbol{O}, \boldsymbol{y}) = P(\boldsymbol{y}|\boldsymbol{O}, \mu_{\mathrm{b}}) P(\boldsymbol{O}|\boldsymbol{s}, \theta) P(\boldsymbol{s}) P(\mu_{\mathrm{b}}) P(\theta)$, where $\theta = \{ \{C^{(\mathrm{p})}[k]\}_{k=1}^{K}, \{C_n^{(\mathrm{a})}\}_{n=0}^{N-1}, \sigma_{\mathrm{p}}^2, \sigma_{\mathrm{a}}^2 \}$ contains all parameters defining the output distributions. The graphical model summarizes the dependency between these variables (Fig. 3).

The conventional inference method [4] uses the EM algorithm by treating $\boldsymbol{s}$ as a latent variable to be marginalized out and θ and μ_{b} as model parameters to be estimated:

- E step: Update $P(s_k = q|\theta, \boldsymbol{O})$ for each frame k and each state q using the Forward-Backward algorithm.
- M step: Update $\boldsymbol{O}$, θ, μ_b using the auxiliary function.

After convergence, MAP estimation for $\boldsymbol{s}$ is performed using the Viterbi algorithm to obtain the final result. We can substitute a point estimation with the Viterbi algorithm for each E step in this algorithm (known as the hard EM algorithm) to accelerate the estimation without loss of accuracy [5].

To improve the inference accuracy, the spectral feature $\boldsymbol{v}$ was incorporated into the HMM in previous research [5]. In this model, the function $P(\boldsymbol{v}|\boldsymbol{s}) P(\boldsymbol{O}|\boldsymbol{s}) P(\boldsymbol{s})$ is maximized instead of $P(\boldsymbol{O}|\boldsymbol{s}) P(\boldsymbol{s})$. This method is designed to detect mora transitions in Japanese speech signals and to restrain the rise of accent commands near mora transition points, but it does not fully utilize linguistic information.

3. PROPOSED METHOD

To further improve the accuracy of estimating F_0 parameters, we propose a new method incorporating linguistic information in the conventional model.

Here, we focus on Japanese as an example of pitch-accent languages. A Japanese utterance is commonly understood as a series of multiple *accentual phrases*. For each accentual phrase, at most one mora called the *accent nucleus* is defined. The accent nucleus in the accentual phrase determines how the accent appears. In more detail, if the first mora in the accentual phrase is the accent nucleus, only the first mora is spoken with a *high* accent, and it is followed by a *low* accent utterance for remaining morae in this phrase. Otherwise, only the utterance between the second mora and and the one defined as the accent nucleus is with a high accent. From the perspective of the Fujisaki model, the accent in each accentual

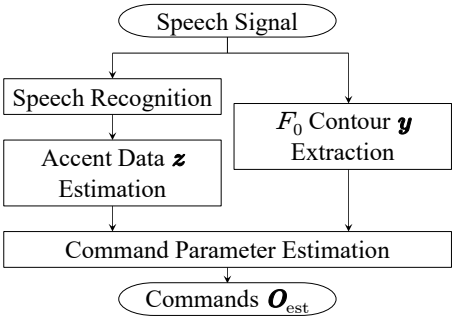

Fig. 4. Block diagram of proposed system.

phrase can be considered to correspond roughly to a single accent command. Therefore, it is expected that the estimation accuracy of the command parameters will be improved by adding accent nucleus information as a clue to restrain the timing of the rise or fall of accent commands.

In the literature, various methods utilizing linguistic information for command estimation have been proposed. A major approach has been deciding initial values of command parameters using linguistic information and iteratively updating them [7, 8]. An alternative approach is first estimating the command parameters by only F_0 information and then modifying them by predetermined procedures using linguistic information [9]. In [10], a method of constructing a statistical model that predicts the command parameters from linguistic information is discussed.

In contrast, here we consider taking advantages of the powerful probabilistic framework for command parameter estimation, by extending the model [5] to incorporate linguistically inferred accent information into the estimation of hidden state sequences. In this approach, both linguistic and prosodic information are incorporated into a single probabilistic model and treated in a unified way.

We introduce a new abstract vector variable representing how "high" the accent is at each frame $\boldsymbol{z} = (z[1], \ldots, z[K])$. $\boldsymbol{z}$ contains linguistic information about accents, and is inferred to be the basic value to determine where F_0 commands occur. Therefore, we introduce probabilistic dependency $P(\boldsymbol{s} \mid \boldsymbol{z})$ and run the estimation algorithm with this new term. This new probability function can be learned using large datasets. When incorporating the term into the EM algorithm, only the E step needs to be modified. When adapting the simplified hard E step, we maximize the function $P(\boldsymbol{O} \mid \boldsymbol{s})P(\boldsymbol{s})P(\boldsymbol{s} \mid \boldsymbol{z})$ with respect to $\boldsymbol{s}$, instead of $P(\boldsymbol{O} \mid \boldsymbol{s})P(\boldsymbol{s})$.

In the case of Japanese, for a given speech script, accent positions are linguistically determined in principle. In our research, we assume that for a given speech signal, each frame is decided to be pronounced with either a high or a low accent. We set each component $z[k]$ to 1 if the k-th frame is pronounced with a high accent and set it to 0 otherwise. As a restriction between $\boldsymbol{z}$ and $\boldsymbol{s}$, we introduce a simple assumption: at frame k that $\boldsymbol{z}$ indicates the point at which the accent is falling($z[k] = 1$ and $z[k+1] = 0$), the accent command must also be falling. In addition, the onsets of the accent commands are only allowed within the frames of $(t - t_0, t + t_1)$, where t is the rising time indicated by $\boldsymbol{z}$. The parameters t_0 and t_1 are predetermined to be $t_0 = 80, t_1 = 40$ [ms].

Compared with the previous method [5] expressing the relationship between prosodic and linguistic information by mediating spectral feature values, the proposed method directly utilizes linguistic information to bind prosodic parameters. From this point of view, this method is expected to achieve further improvement in accuracy.

4. EXPERIMENTS

We built an end-to-end command parameter estimation system to evaluate its accuracy on real speech data (Fig. 4).

For a given speech signal, This system first performs speech recognition to acquire linguistic information. Here, we used Julius [11]. Phoneme alignment was also performed. Next, accentual phrases and the position of the accent kernel for each accentual phrase were estimated using TASET [12], and $\boldsymbol{z}$ was calculated. Then, we extracted F_0 patterns $\boldsymbol{y}$ from the speech signal by the method [13], and initialized $\boldsymbol{O}$ using Narusawa's method [6]. Finally, we iterated the EM algorithm 20 times to get the estimated parameter values. We also conducted another experiment using human-annotated accent kernel information to evaluate the upper limit of the performance of our model when linguistic analysis is ideal.

The speech data was excerpted from the ATR Japanese sentence database B-set [14]. We used 503 sentences spoken by one male speaker (MHT), of which the first 200 sentences were used for learning. The baseline values μ_{b} were fixed to the minimum value of F_0 in the voiced segments.

Throughout the experiments, we used the values $\alpha = 3.0$ rad/s, $\beta = 20.0$ rad/s, $N = 10, \sigma_{\mathrm{p}}^2 = \sigma_{\mathrm{a}}^2 = 0.03^2$, and set the time shift per frame to 8 ms.

To measure the estimation accuracy, we used the deletion and insertion rates of estimated commands. These are defined as follows. First, we match the estimated and ground truth command sequences on a command-by-command basis with the dynamic programming algorithm. By using a predefined time difference tolerance S, the estimated commands that have a match are judged to be "matched." In the case of accent commands, we compare the mean value of the onset time difference and the offset time difference with S. Let N, N_{est}, and N_{match} be the number of commands in the ground truth data, in the estimated result, and judged to be matched, respectively. The deletion/insertion rates are then defined as $p_{\mathrm{del}} = (N - N_{\mathrm{match}})/N$ and $p_{\mathrm{ins}} = (N_{\mathrm{est}} - N_{\mathrm{match}})/N$. In addition, we call the value $2p_{\mathrm{del}}p_{\mathrm{ins}}/(p_{\mathrm{del}} + p_{\mathrm{ins}})$as 'F-measure' for convenience.

We fixed $S = 0.1$ [sec], which is approximately a typical duration of one mora in Japanese utterances. These criteria do not concern the magnitudes of the commands because we

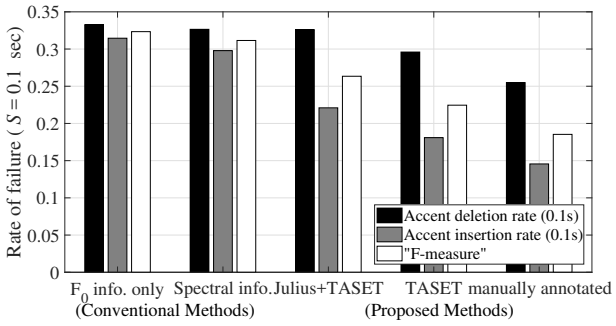

Fig. 5. Comparison of deletion/insertion rates. From the left, each column represents the method using only F_0, the conventional method using spectral feature values [5], and the proposed methods, respectively. The proposed methods were conducted under three conditions: fully automatic linguistic analysis (Julius+TASET), annotated labels substituting speech recognition (TASET), and all hand-labeled linguistic analysis (manually annotated).

Table 1. Average computation time per one sentence.

Process [s]	Conventional	Proposed
Speech recognition	-	0.480
Accent kernel estimation	-	0.514
EM algorithm (20 iterations)	2.411	4.216
Total estimation	2.411	5.210
Average duration of signal	5.279	

are interested in the appropriateness of estimation in terms of the number of correctly estimated commands.

Fig. 5 shows the insertion and deletion rates of command functions estimated with the conventional and proposed methods. Incorporating automatically obtained linguistic information by using Julius and TASET reduced the 'F-measures' by 19%. Using manually annotated ground-truth linguistic information reduced the value by 43%. Although the assumptions of the proposed model are very simple, this is a significant improvement when compared to the value of 4% obtained by the previous method [5] that incorporated spectral features to detect mora transitions. This fact shows that use of linguistic information is quite effective for improvement of the accuracy of the conventional stochastic estimation method. It also suggests that it is worth considering more complicated stochastic models which describe the relationship between linguistic and prosodic information for further improvement of the accuracy.

The computation times spent for the estimation are summarized in Table 1. We conducted the experiment using a desktop computer with Intel Core i7 6700K CPU (programs are not parallelized) and NVIDIA Quadro K620 GPU (only for DNN-based speech recognition). We compared the proposed method with the conventional one [5] which uses only F_0 information. This table shows that the proposed method

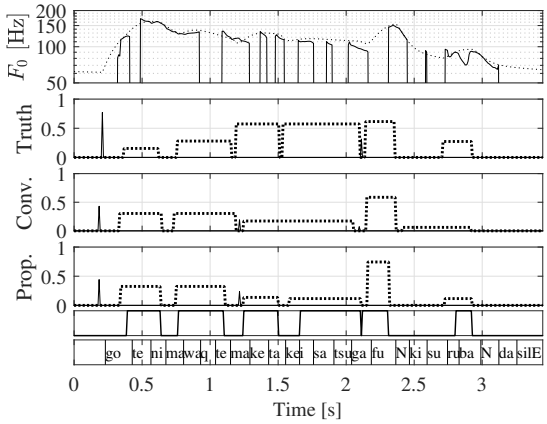

Fig. 6. Example of inferring results for one Japanese sentence (J27:*Goteni mawatte maketa keisatsuga funkisuru banda*, It's the police's turn to show their fighting spirit who once fell behind and lost). The top row shows the F_0 contours. The following rows show the ground-truth and inferred command parameters, by the method using only F_0 and the method executing the linguistic process automatically. The solid lines and dotted lines represent phrase and accent commands, respectively. The next row shows the estimated accent included in TASET output. The bottom row shows the mora segmentation data.

works fast enough for real-time operation even when considering the contribution of linguistic information processing.

Fig. 6 shows an example of the estimated results. It can be seen that the proposed method outputs more reasonable command parameters, and that the false accent command estimations of the conventional method are rectified.

5. CONCLUSION

We described a novel method to estimate F_0 command parameters. Compared to the existing probabilistic models for this task, the advantage of the proposed method is its use of linguistic information obtained from the input speech. We showed that it significantly reduced the command estimation errors. While we focused on Japanese language in this paper, our method is expected to be effective also on other languages as long as linguistic information and prosody are correlated to each other. Future work will include constructing probabilistic models reflecting the relationships between linguistic information and F_0 commands more precisely. Integrating some other sources of information that may further improve the command estimation accuracy is also an important task.

6. ACKNOWLEDGEMENTS

We thank Prof. Hiroshi Saruwatari and Dr. Shinnosuke Takamichi for their helpful comments and support, and Prof. Keikichi Hirose for providing us with the manually annotated F_0 command data associated with the ATR speech database.

7. REFERENCES

[1] Keikichi Hirose and Jianhua Tao, *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis*, Springer, 2015.

[2] Hiroya Fujisaki, "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour," *Vocal physiology: Voice production, mechanisms and functions*, pp. 347–355, 1988.

[3] H. Mixdorff, "A novel approach to the fully automatic extraction of fujisaki model parameters," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, 2000, vol. 3, pp. 1281–1284 vol.3.

[4] Hirokazu Kameoka, Kota Yoshizato, Tatsuma Ishihara, Kento Kadowaki, Yasunori Ohishi, and Kunio Kashino, "Generative modeling of voice fundamental frequency contours," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1042–1053, 2015.

[5] Ryotaro Sato, Hirokazu Kameoka, and Kunio Kashino, "Fast algorithm for statistical phrase/accent command estimation based on generative model incorporating spectral features," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5595–5599.

[6] Shuichi Narusawa, Nobuaki Minematsu, Keikichi Hirose, and Hiroya Fujisaki, "A method for automatic extraction of model parameters from fundamental frequency contours of speech," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. IEEE, 2002, vol. 1, pp. I–509.

[7] Hiroya Fujisaki and Sumio Ohno, "Prosodic parameterization of spoken japanese based on a model of the generation process of f_0 contours," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*. IEEE, 1996, vol. 4, pp. 2439–2442.

[8] Hiromasa Ogawa and Yoshinori Sagisaka, "Automatic extraction of f0 control parameters using utterance information," in *Speech Prosody 2004, International Conference*, 2004.

[9] Keikichi Hirose, Yusuke Furuyama, Shuichi Narusawa, Nobuaki Minematsu, and Hiroya Fujisaki, "Use of linguistic information for automatic extraction of f_0 contour generation process model parameters," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[10] Toshio Hirai, Naoto Iwahashi, Norio Higuchi, and Yoshinori Sagisaka, "Automatic extraction of f_0 control rules using statistical analysis," in *Progress in speech synthesis*, Jan PH Van Santen, Richard Sproat, Joseph Olive, and Julia Hirschberg, Eds., chapter 28, pp. 333–346. Springer Science & Business Media, 2013.

[11] Akinobu Lee and Tatsuya Kawahara, "Recent development of open-source speech recognition engine julius," in *Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference*. Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, International Organizing Committee, 2009, pp. 131–137.

[12] Masayuki Suzuki, Ryo Kuroiwa, Keisuke Innami, Shumpei Kobayashi, Shinya Shimizu, Nobuaki Minematsu, and Keikichi Hirose, "Accent sandhi estimation of tokyo dialect of japanese using conditional random fields," *IEICE TRANSACTIONS on Information and Systems*, vol. 100, no. 4, pp. 655–661, 2017.

[13] Tomohiro Nakatani, Shigeaki Amano, Toshio Irino, Kentaro Ishizuka, and Tadahisa Kondo, "A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments," *Speech Communication*, vol. 50, no. 3, pp. 203–214, 2008.

[14] Akira Kurematsu, Kazuya Takeda, Yoshinori Sagisaka, Shigeru Katagiri, Hisao Kuwabara, and Kiyohiro Shikano, "Atr japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.