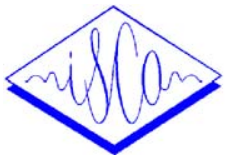


EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

RESTORATION OF PITCH PATTERN OF SPEECH BASED ON A PITCH GENERATION MODEL

Hiroshi Shimodaira , *Mitsuru Nakai* and *Akihiro Kumata*

School of Information Science,
Japan Advanced Institute of Science and Technology,
Tatsunokuchi, Ishikawa, 923-12 Japan
E-mail: sim@jaist.ac.jp

ABSTRACT

In this paper a model-based approach for restoring a continuous fundamental frequency (F_0) contour from the noisy output of an F_0 extractor is investigated. In contrast to the conventional pitch trackers based on numerical curve-fitting, the proposed method employs a quantitative pitch generation model, which is often used for synthesizing F_0 contour from prosodic event commands for estimating continuous F_0 pattern. An inverse filtering technique is introduced for obtaining the initial candidates of the prosodic commands. In order to find the optimal command sequence from the commands efficiently, a beam-search algorithm and an N-best technique are employed. Preliminary experiments for a male speaker of the ATR B-set database showed promising results both in quality of the restored pattern and estimation of the prosodic events.

1. INTRODUCTION

Pitch contours are well known as a medium for representing the most significant part of prosodic information. Although it is largely expected that pitch contours can be effectively used for automatic speech recognition, no method has been developed for detecting pitch frequencies with high accuracy. Moreover, even if we assume that the detected pitch frequency is reliable, due to undefined pitch frequency for unvoiced sounds of a speech; the continuous pitch contours of an utterance can not be obtained. Therefore, some sort of method for smoothing the pitch contours is necessary to detect prosodic information from pitch contours. Most of the smoothing methods widely used are based on numerical curve fitting algorithms such as linear line fitting, spline curve fitting and so on. However, these methods are likely to lose prosodic cues which were contained in the original pitch contours. In the present approach, this problem has been extensively investigated and a method has been proposed for pitch contour restoration based on a super-positional prosodic model.

The Fujisaki model [1] is used as the prosodic model of the current study. In this model, F_0 which is a function of time t is given by

$$\ln F_0(t) = \ln F_{\min} + \sum_{i=1}^{I} A_{p_i} G_p(t - T_{pi}) + \sum_{j=1}^{J} A_{a_j} \{G_a(t - T_{aj}) - G_a(t - (T_{aj} + \tau_{aj}))\}, \quad (1)$$

where

$$G_p(t) = \begin{cases} \alpha^2 t e^{-\alpha t}, & (t \geq 0), \\ 0, & (\text{otherwise}), \end{cases} \quad (2)$$

indicates the impulse response function of the phrase control mechanism, and

$$G_a(t) = \begin{cases} \min[1 - (1 + \beta t)e^{-\beta t}, \theta], & (t \geq 0), \\ 0, & (\text{otherwise}), \end{cases} \quad (3)$$

indicates the step response function of the accent control mechanism.

The impulse and step signals driving the model are called "phrase commands", "accent commands" respectively and they are often called "prosodic events" or "commands" as a whole. Compared to the original input contours, if we can choose a suitable set of prosodic events from the given raw F_0 contours, then the F_0 contour restored by the model will have better characteristics for further prosodic analysis for speech recognition.

Although this model has been widely used in the area of speech synthesis, it is still an open problem to develop a method for automatic estimation of the reasonable set of commands from a given F_0 contour [2, 3]. Here, we propose a new method that employs inverse filtering technique to find the initial candidates of the prosodic commands, and it selects the best fit set of commands for the input F_0 contours.

2. OUTLINE

Fig. 1 shows an outline of the proposed method. At first, raw F_0 contours are obtained from a pitch extractor. The input to the extractor is the

frame-by-frame input speech signal sequence and the output of the same is fed into two inverse filters that generate candidates for accent and phrase commands. Since the inverse filters produce a lot

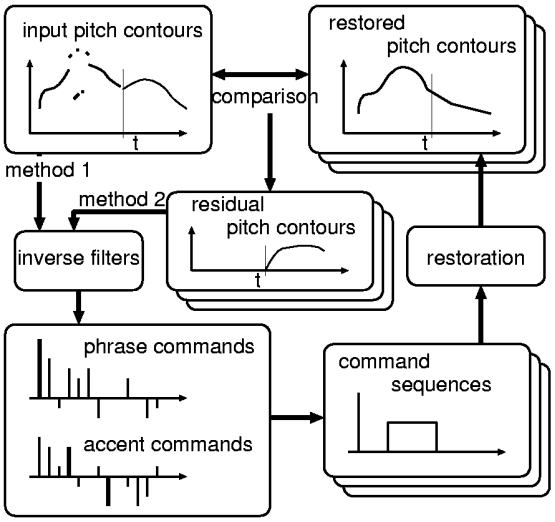


Figure 1: Outline of the system

of noisy output against the raw F_0 contours with enormous discontinuities, a mechanism for finding the plausible prosodic commands from the filter outputs is required. This is implemented by combining a beam-search algorithm which works synchronously with time and an N-best search algorithm for selecting the possible candidate sequences. In this algorithm, the candidates are ranked according to their fitness to the input F_0 contours. The fitness is measured by a scale of squared error (distortion) between the original input F_0 contours and the restored candidate sequence.

3. RESTORATION ALGORITHM

3.1. Inverse Filtering

Although the F_0 generating mechanism of the Fujisaki model is simple, estimation of commands for generating a given F_0 contour is quite hard and not straightforward. This is due to the fact that the estimation of the input of the model is an ill-conditioned inverse problem. Hence, successive approximation algorithm which is sometimes called "Analysis-by-Synthesis" (A-b-S) is widely used to estimate the commands. On the other hand, Fujisaki and his colleagues proposed a rather straightforward approach using an inverse filter technique [4]. In their approach, two types of inverse filters were designed, one is for detecting phrase commands and the other is for detecting accent commands. In the present study, the same filters are employed.

The transfer function of the filter for detecting phrase commands and accent commands are denoted by $H_p^{-1}(z)$ and $H_a^{-1}(z)$ respectively, and they are given by

$$H_p^{-1}(z) = \frac{z^{-1} - 2 - 2(\alpha T + 1)z^{-1} + (\alpha T + 1)^2}{\alpha^2 T^2} \quad (4)$$

$$H_a^{-1}(z) = \beta^{-2} T^{-3} \left\{ (\beta T - 1)^2 - (\beta T + 1)(\beta T + 3) z^{-1} + (2\beta T + 3) z^{-1} - 2 - z^{-1} - 3 \right\}. \quad (5)$$

These filters were designed under the assumption that at a certain time instance, the output of the Fujisaki model will get dominated by only one command of the command sequence. Therefore, the filters would work reasonably only under the conditions that the previously mentioned assumption holds, and the observed F_0 contours are continuous and noise free.

Fig. 2 shows an example of applying the inverse filters to the observed F_0 pattern of a real speech. It can be seen that due to the noisy outputs of the filters, it is not easy to find the true command sequence by employing a simple decision logic.

The problem related to the super-positional effect of the commands could be solved by feeding the modified F_0 pattern to the filters for eliminating the influence of the preceding commands from the pattern.

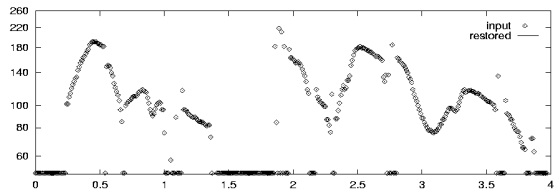
Now, if we assume that the correct command sequence up to processing time t_p is estimated, and represent the observed pitch pattern in logarithmic scale and the generated pitch patterns as $P(t)$ and $\tilde{P}_{t_p}(t)$ respectively, then the residual pattern expressed by $R_{t_p}(t) = P(t) - \tilde{P}_{t_p}(t)$ will have no components corresponding to the commands occurred before t_p. This $R_{t_p}(t)$ could be effectively used as an input to the filters for capacitating the same to work properly. It is not possible to determine the correct command sequence for a certain time interval, and therefore, this residual pattern is to be calculated for every hypotheses of the command sequence at time t_p. Namely, for the m-th hypothesis,

$$R_{m,t_p}(t) = P(t) - \tilde{P}_{m,t_p}(t), \quad (6)$$

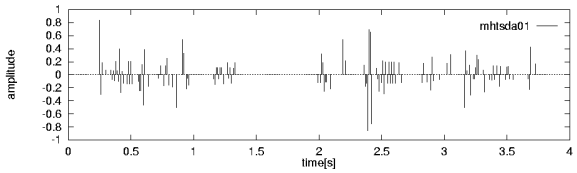
where $\tilde{P}_{m,t_p}(t)$ is the generated F_0 value in logarithmic scale at time t ($t \leq t_p$). In contrast to "method-1" in which the observed F_0 contour is fed directly to the filters (see Fig. 1.), the proposed compensation method will be called as "method-2".

3.2. Distortion Measure

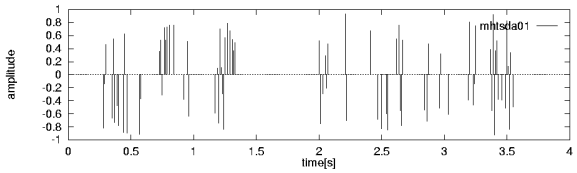
As mentioned in the previous section, it is necessary to find a mechanism for selecting the proper commands from the noisy output of the inverse filters. In order to evaluate the quality of the chosen command sequence, a distortion measure is employed. However, a straightforward definition of the measure in the domain of command sequence is very complex, and therefore, in the present approach it is defined in



(a) Observed F_0 contours (input to the filters)



(b) Output of the phrase command detector



(c) Output of the accent command detector

Figure 2: Example of inverse filtering

the domain of F_0 contours as a distortion of the regenerated F_0 contour from the command sequence against the observed F_0 contour. For the m-th candidate command sequence at processing time instance t_p, if D_{m,t_p} is the cumulated distortion of $\tilde{P}_{m,t_p}(t)$ up to processing time t_p against the observed $\log F_0$ value $P(t)$, then D_{m,t_p} can be defined as

$$D_{m,t_p} = \sum_{t=0}^{t_p} w(t)(P(t) - \tilde{P}_{m,t_p}(t))^2, \qquad (7)$$

where the weight function $w(t)$ is given by the pitch extraction reliability measure with its value in the interval of $[0, 1]$.

3.3. Search algorithm

Since a large number of possible combinations of command sequences can be generated from the output of the inverse filters, it is not feasible to enumerate all of the command sequences and calculate their distortion (7). Hence, in the present approach; an efficient searching algorithm is developed for choosing the plausible candidate sequences. The algorithm is combined with a beam-search algorithm which works synchronously with time and an N-best search algorithm (Fig. 3).

In the proposed algorithm, the following steps are taken according to the processing time t_p which starts at 0 and extends until the ending frame of the utterance is reached.

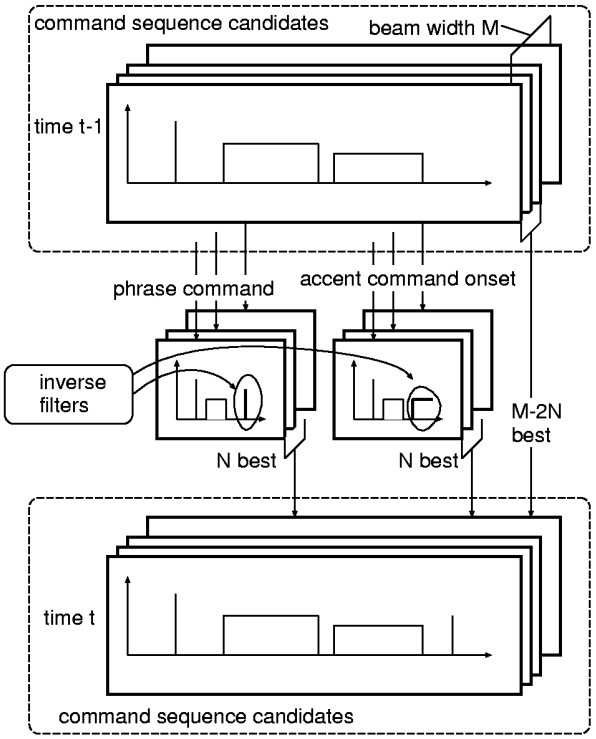


Figure 3: Search algorithm

1. Calculate the distortion D_{m,t_p} for the top M candidates.

2. Select the candidates based on the following rules.

   If a phrase command A_p is detected at t_p then choose the top N candidates according to the distortion D_{m,t_p} among the M candidates at time $t_p - 1$ that can be followed by the phrase command A_p. Add the phrase command A_p to the end of each candidate sequence.

   If an accent command A_a is detected at t_p then choose the top N candidates according to the distortion D_{m,t_p} among the M candidates at time $t_p - 1$ that can be followed by the accent command A_a. Add the accent command A_a to the end of each candidate sequence.

3. Choose top $M - 2N$ candidates among the M candidates at $t_p - 1$ assuming that no command occurred at time t_p.

4. Experiments

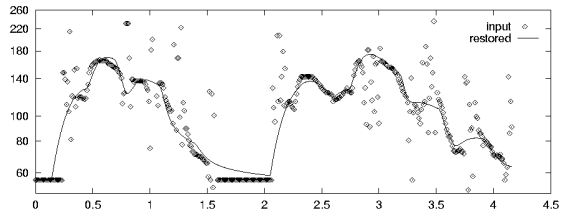
4.1. Experimental conditions

For evaluating the performance of the proposed method, 50 sentences were chosen from 503 sentences of the ATR continuous speech database (set-B)

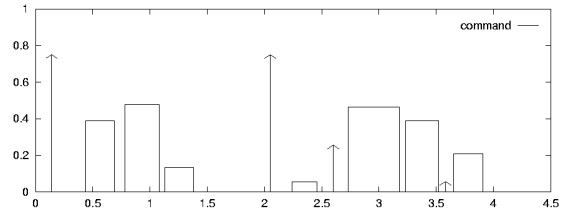
uttered by a male speaker MYI. As a pitch determinant algorithm, the "Lag-window method" [5] was employed with analysis frame interval of 10ms. As for the parameters of the inverse filters, α was set to 3.0 and β was set to 20.0.

4.2. Results

Fig. 4 shows an example of pitch restoration by employing method-2. In Fig. 4, it can be seen that the proposed method succeeded in capturing most of the prosodic events and the restored pitch pattern appears to be reasonable.



(a) Original F_0 and restored F_0 pattern



(b) Estimated command sequence

Figure 4: Example of restoration by method-2

Fig. 5 shows the resultant performance measured in distortion against the search parameters (the beam-width M and N-best). It is clear from the figure that method-2 works better than method-1. However, there exists some tradeoff between the parameters N and M.

Another comparison has been done in the basis of the distortion against the ideal F_0 contours which were generated from the Fujisaki model where the model parameters were determined by A-b-S with human guide. In Table. 1, it can be seen that compared to the the numerical curve smoothing technique such as liner-line fitting and moving average filter, the proposed method-2 gives closer F_0 patterns to the ideal patterns. However, if command sequence estimation is taken into account, then the present result is not close to the human aided A-b-S results. The correct rate of command estimation was around 30% with $\pm$ 100ms discrepancy from the A-b-S command location.

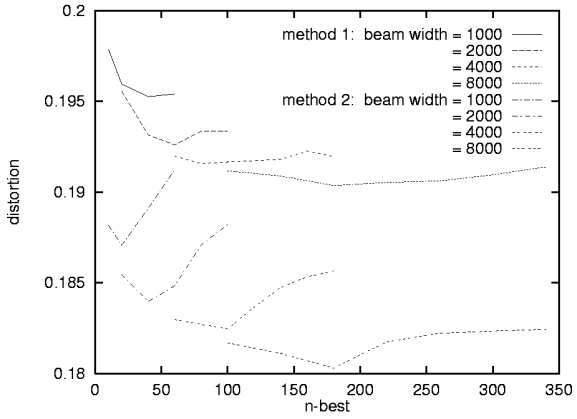


Figure 5: Distortion against search widths

smoothing method	distortion (log Hz/frame)
(without smoothing)	0.332
linear-line fitting	0.259
moving average filter	0.216
method-1	0.206
method-2	0.192

Table 1: Distortion against ideal F_0 pattern

5. CONCLUSION

Although it is still an open problem to establish an algorithm for estimating the command sequence of the Fujisaki model, the proposed method based on the inverse-filtering technique seems promising. The proposed method is to be improved farther by designing filters that are robust against noisy input and by adding some restrictions for choosing the commands.

REFERENCES

[1] K. Hirose and H. Fujisaki. Analysis and synthesis of voice fundamental frequency contours of spoken sentences. *ICASSP-82*, 2:950–953, 1982.

[2] Sumio Ohno Hiroya Fujisaki. Prosodic parameterization of spoken japanese based on a model of the generation process of F_0 contours. *ICSLP 96*, 4:2439–2442, September 1996.

[3] Edouard Geoffrois. A pitch contour analysis guided by prosodic event detection. *Eurospeech'93*, 2:793–796, September 1993.

[4] Hiroya Fujisaki, Sumio Ohno, and Yutaka Wada. A method for automatic estimation of parameters of a moel for the generation process of fundamental frequency contours of speech. *Spring Meeting of Acousti Society of Japan*, pages 17–18, March 1995.

[5] Hiroshi Shimodaira and Masayuki Kimura. Accent phrase segmentation using pitch pattern clustering. *ICASSP-92*, 1:217–220, March 1992.