# AUTOMATIC IDENTIFICATION FOR SINGING STYLE BASED ON SUNG MELODIC CONTOUR CHARACTERIZED IN PHASE PLANE

**Tatsuya Kako†, Yasunori Ohishi‡, Hirokazu Kameoka‡, Kunio Kashino‡, Kazuya Takeda†**

†Graduate School of Information Science, Nagoya University

‡NTT Communication Science Laboratories, NTT Corporation

kako@sp.m.is.nagoya-u.ac.jp, ohishi@cs.brl.ntt.co.jp, kameoka@eye.brl.ntt.co.jp
kunio@eye.brl.ntt.co.jp, kazuya.takeda@nagoya-u.jp

## ABSTRACT

A stochastic representation of singing styles is proposed. The dynamic property of melodic contour, i.e., fundamental frequency ($F_0$) sequence, is assumed to be the main cue for singing styles because it can characterize such typical ornamentations as *vibrato* . $F_0$ signal trajectories in the phase plane are used as the basic representation. By fitting Gaussian mixture models to the observed $F_0$ trajectories in the phase plane, a parametric representation is obtained by a set of GMM parameters. The effectiveness of our proposed method is confirmed through experimental evaluation where 94.1% accuracy for singer-class discrimination was obtained.

## 1. INTRODUCTION

Although no firm definition has yet been established for "singing style" in musical information processing research, several studies have reported the relationship between singing styles and such signal features as singing formant [1, 2] and singing ornamentations. Various research efforts have been made to characterize ornamentations by the acoustical property of the sung melody, i.e., *vibrato* [3–11], overshoot [12], and fine fluctuation [13]. The importance of such melodic features for perceiving singer individuality was also reported in [14] based on psycho-acoustic experiments. They concluded that the average spectrum and the dynamical property of the $F_0$ sequence affect the perception of the individuality. Those studies suggest that singing style is related to the local dynamics of a sung melody that does not contain any musical information. Therefore, in this study, we focus on the local dynamics of the $F_0$ sequence, i.e., the melodic contour, as a cue of singing style and propose a parametric representation as a model for singing styles.

On the other hand, very few application systems have been reported that use the local dynamics of a sung melody. [15] reported a singer recognition experiment using *vibrato* . [16] reported a method for evaluating singing skill through the spectrum analysis of the $F_0$ contour. Although these studies try to use the local dynamics of melodic contour as a cue for ornamentation, no systematic method has been proposed for characterizing singing styles. A lag system model for typical ornamentations was reported in [14, 17–19]; however, variation of singing styles was not discussed.

In this paper, we propose a stochastic phase plane as a graphical representation of singing styles and show its effectiveness for singing style discrimination. One merit of this representation to characterize singing style is that since neither an explicit detection function for ornamentation like *vibrato* nor estimation of the target note is required, it is robust to sung melodies.

In a previous paper [20], we applied this graphical representation of the $F_0$ contour in the phase plane to a query-by-hamming system and neutralized the local dynamics of the $F_0$ sequence so that only musical information was utilized for the query. In contrast, in this study, we use the local dynamics of the $F_0$ sequence for modeling singing styles and disregard the musical information because musical information and singing style are in a dual relation.

In this paper, we also evaluate the proposed representation through a singer-class discrimination experiment in which we show that our proposed model can extract the dynamic properties of sung melodies shared by a group of singers.

In the next section, we propose stochastic phase plane (SPP) as a stochastic representation of the melodic contour and show how singing ornamentations are modeled by the proposed SPP. In Section 3, we experimentally show the effectiveness of our proposed method through singer class discrimination experiments. Section 4 discusses the obtained results and concludes this paper.

## 2. STOCHASTIC REPRESENTATION OF THE DYNAMICAL PROPERTY OF MELODIC CONTOUR

### 2.1 $F_0$ signal in the Phase Plane

Such ornamental expressions in singing as *vibrato* are characterized by the dynamical property of their $F_0$ signal. Since the $F_0$ signal is a controlled output of the human speech production system, its basic dynamical characteristics can be related to a differential equation. Therefore, we can use the phase plane, which is the joint plot of a variable and its time derivative, i.e., $(x, \dot{x})$, to depict its dynamical property.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

© 2009 International Society for Music Information Retrieval.

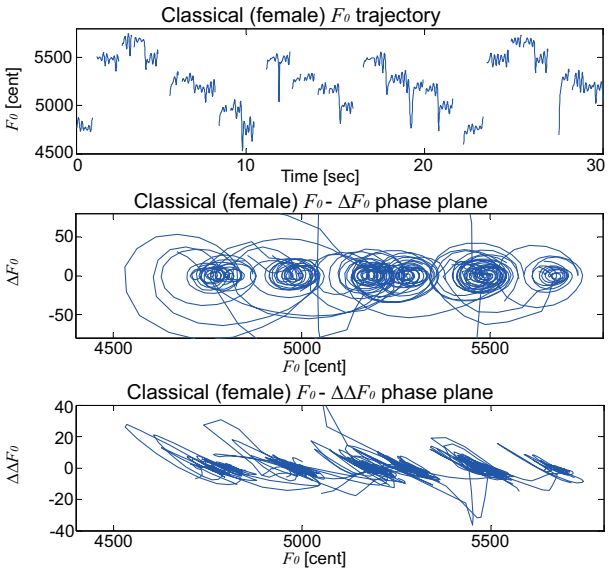


**Figure 1**. Melodic contour (top) and corresponding phase planes for $F_0$-$\Delta F_0$ (middle) and $F_0$-$\Delta\Delta F_0$ (bottom)

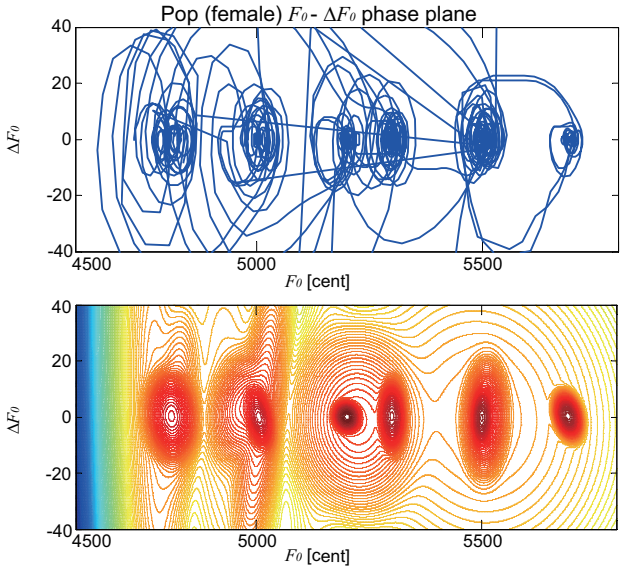


**Figure 2**. Gaussian mixture model fitted to $F_0$ contour in phase plane

Although the signal sequence is not given as an explicit function of time, $F_0(t)$, but as a sequence of numbers, $\{F_0(n)\}_{n=1,\cdots,N}$, we can estimate the time derivative using the *delta*- coefficient given by

$$\Delta F_0(n) = \frac{\sum_{k=-K}^{K} k \cdot F_0(n+k)}{\sum_{k=-K}^{K} k^2}, \quad (1)$$

where $2K$ is the window length for calculating the dynamics. Changing the window length extracts different aspects of the signal property.

An example of such a plot for a given melodic contour is shown in Fig. 1. Here, the $F_0$ signal (top), the phase plane (middle), and the second order phase plane, which is given by the joint plot of $F_0$ and $\Delta\Delta F_0$ (bottom), are plotted. The singing ornamentations are depicted as the local behavior of the trajectory around the centroids that commonly represent target musical notes. *Vibrato* in singing, for example, is shown as circular trajectories centered at target notes. In the second order plane, the trajectories appear as lines with a slope of -45 degrees. This shows that the relationship between $F_0$ and $\Delta\Delta F_0$ is given as

$$\Delta\Delta F_0 = -F_0. \quad (2)$$

Hence, the sinusoidal component is imposed in the given signal. Over/under-shoots to the target note are represented as spiral patterns around the note.

### 2.2 Stochastic representation of Phase Plane

Once a singing style is represented as a phase plane trajectory, parameterizing the representation becomes an issue for further engineering applications. Since the $F_0$ signal is not deterministic, i.e., it varies across singing behaviors, a stochastic model must be defined for the parameterization. By fitting a parametric probability density function to the trajectories in the phase plane, we can build a stochastic phase plane (SPP) and use it for characterizing the melodic contour. A common feature of the trajectories in the phase plane is that most of their segments are distributed around the target note, and therefore the distribution's histogram is multimodal, but each mode can be represented by a simple symmetric 2d or 3d-pdf. Therefore, Gaussian mixture model (GMM),

$$\sum_{m=1}^{M} \lambda_m \mathcal{N}(\mathbf{f}_0(n); \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \quad (3)$$

where

$$\mathbf{f}_0(n) = [F_0(n), \Delta F_0(n), \Delta\Delta F_0(n)]^{\mathrm{T}}, \quad (4)$$

is adopted for the modeling. $\mathcal{N}(\cdot)$ is a Gaussian distribution, and

$$\Theta = \{\lambda_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}_{m=1,\cdots,M}, \quad (5)$$

are parameters of the model, each of which represents the relative frequency, the mean vector, and the covariance matrix of each Gaussian.

A GMM trained for $F_0$ contours in the phase plane is depicted in Fig. 2. A smooth surface is trained through model fitting. The horizontal deviations of each Gaussian represent the stability of the melodic contour around the target note, but the vertical deviations represent the *vibrato* depth. In this manner, singing styles can be modeled by set of parameters $\Theta$ of the stochastic phase plane.

### 2.3 Examples of Stochastic Phase Plane

In Fig. 3, the $F_0$ signals of three female singers are plotted: professional classical, professional pop, and an amateur. A deep *vibrato* is observed as a large vertical deviation in the Gaussians in the professional classical singer's plot. On the other hand, the amateur's plot is characterized by large horizontal deviations. Although deep *vibrato* is not observed in the plot for the professional pop singer, its smaller horizontal deviation shows that she accurately sang the melody.

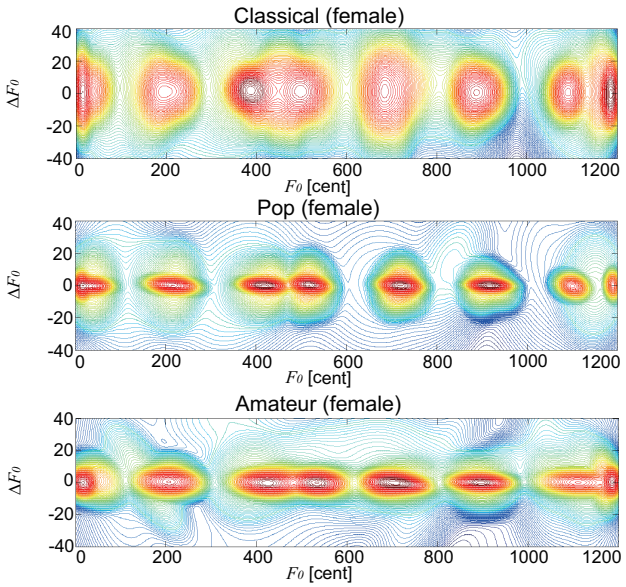


**Figure 3**. Stochastic phase plane models for professional classical (top), professional pop (middle), and amateur (bottom)

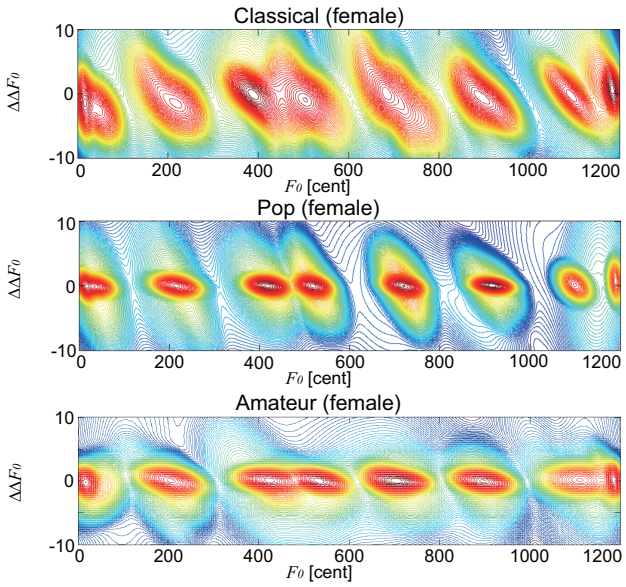


**Figure 4**. 2nd order stochastic phase plane models for professional classical (top), professional pop (middle), and amateur (bottom)

**Table 1**. Signal analysis conditions for $F_0$ estimation. Harmonical PSD pattern matching [21] is used with these parameters.

| | |
|---|---|
| Signal sampling freq. | 16 kHz |
| $F_0$ estimation window length | 64 ms |
| Window function | Hanning window |
| Window shift | 10 ms |
| $F_0$ contour smoothing | 50 ms MA filter |
| $\Delta$ coefficient calculation | $K = 2$ |

The stochastic representations of the second order phase plane are also shown in Fig. 4. Strong negative correlations between $F_0$ and $\Delta\Delta F_0$ can be found only in the plot for the professional classical singer that also indicates deep *vibrato* in the singing style.

## 3. EXPERIMENTAL EVALUATION

The effectiveness of using SPP to discriminate different singing styles is evaluated experimentally.

### 3.1 Experimental set up

The following singing signals of six singers were used: one of each gender in the categories of professional classical, professional pop, and amateur. With/without musical accompaniment, each subject sang songs with Japanese lyrics and hummed. The songs were "*Twinkle, Twinkle, Little Star*", and "*Ode to Joy*" and five etudes. A total of 102 song signals was recorded.

The $F_0$ contour was estimated using [21]. The signal processing conditions for calculating $F_0$, $\Delta F_0$, and the $\Delta\Delta F_0$ contours are listed in Table 1.

Since the absolute pitch of the song signals differ across singers, we normalized them so that only the singing style of each singer is used in the experiment. Normalization was done in the procedure below. First, the $F_0$ frequency in `[Hz]` is converted to `[cent]` by

$$1200 \times \log_2 \frac{F_0}{440 \times 2^{3/12-5}} \quad \texttt{[cent]}. \tag{6}$$

Then the local deviations from the tempered clavier are calculated by the residue operation mod$(\cdot)$:

$$\mathrm{mod}\ (F_0 + 50, 100). \tag{7}$$

Obviously, after this conversion, the $F_0$ value is limited to $(0, 100)$ in `[cent]`.

### 3.2 Discrimination Experiment

The discrimination of three singer classes, i.e., professional classical, professional pop, and amateur, was performed based on the maximum *a posteriori* probability (MAP) decision:

$$\begin{aligned} \hat{s} &= \arg\max_s \left[p(s|\{F_0, \Delta F_0, \Delta\Delta F_0\})\right] \\ &= \arg\max_s \left[\frac{1}{N}\sum_{n=1}^{N} \log p(\mathbf{f}_0(n)|\Theta_s) + \log p(s)\right] \end{aligned} \tag{8}$$

where $s$ is the singer-class id and $\Theta_s$ is the model parameters of the $s^{th}$ singer-class. We used "*Twinkle-Twinkle, Little Star*" and five etudes sung by singers from each singer class for training and "*Ode to Joy*" sung by the same singers for testing. Therefore the results are independent from sung melodies but closed in singers. $N$ is the length of the signal in the samples. Since we assumed an equal *a priori* probability for singer-class distribution $p(s)$, the above MAP decision is equivalent to the Maximum Likelihood decision.

### 3.3 Results

Fig. 5 shows the accuracy of the singer-class discrimination. The best is attained for a 13-second input sig-

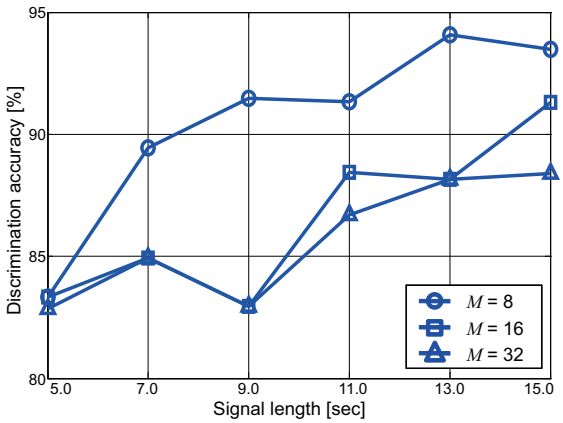


**Figure 5**. Accuracy in discriminating three singer classes

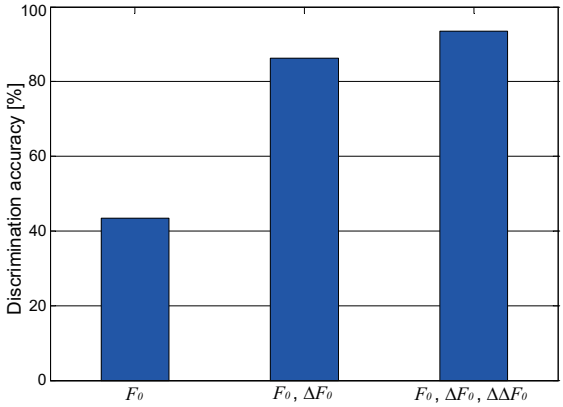


**Figure 6**. Comparing accuracy in discriminating singer classes

nal. The accuracy increases with the length of the test signal and 94.1% is attained with an 8-mixture GMM for singer-class models, when a 13-second signal is available for the test input. No significant improvement in accuracy was found for the longer test input because more song-dependent information contaminated the test signal. Fig. 6 compares the accuracy of singer-class discriminations using the three sets of features: $F_0$ only, ($F_0$, $\Delta F_0$), and ($F_0$, $\Delta F_0$, $\Delta\Delta F_0$). As shown in the figure, by combining $F_0$ and $\Delta F_0$, the discrimination error rate becomes half of the error when only using $F_0$. Combining second order derivative $\Delta\Delta F_0$ further reduces the error but not as much as the case of $\Delta F_0$. These results show that the proposed stochastic representation of the phase plane effectively characterizes the singing styles of the three singer classes.

## 4. DISCUSSION

Our proposed method for representing and parameterizing the $F_0$ contour effectively discriminates the three typical singer classes, i.e., professional classical and pop, and amateurs. To confirm that the method models the singing styles (and not singer individuality), we compared our proposed representation with MFCC under two conditions. As a closed condition, we trained three MFCC-GMMs using "*Twinkle-Twinkle, Little Star*" and five etudes sung by six (male and female professional classic, professional pop, and amateur) singers and used "*Ode to Joy*" sung by

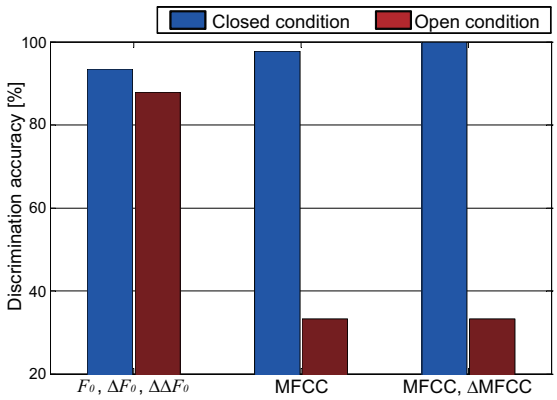


**Figure 7**. Comparing proposed representation with MFCC under two conditions

the same singers for testing. On the other hand, as an open condition, we evaluated the MFCC-GMMs through a singer independent manner where singer-class models (GMMs) were trained by female singer data and tested by male singer data. As shown in Fig. 7, the performances of the MFCC-GMM and the proposed method are almost identical (95.0%) in the closed condition. However, in the new (unseen) singer experiment, the result of the MFCC-GMM system significantly degraded to 33.3%, but the proposed method attained 87.9% accuracy. These results suggest that the MFCC-GMM system does not model the singing style but discriminates singer individuality. However, since SPP-GMM can correctly classify even an unseen singer's data, our proposed representation models the $F_0$ dynamic characteristics common within a singer class better than singer individuality.

## 5. SUMMARY

In this paper, we proposed a model for singing styles based on the stochastic graphical representation of the local dynamical property of the $F_0$ sequence. Since various singing ornamentations are related to signal production systems described by differential equations, phase plane is a reasonable space for depicting singing styles. Furthermore, the Gaussian mixture model effectively parameterizes the graphical representation; therefore, more than 90% accuracy can be achieved in discriminating the three classes of singers.

Since the scale of the experiments was small, increasing the number of singers and singer classes is critical future work. Evaluating the robustness of the proposed method to noisy $F_0$ sequences estimated under such realistic singing conditions as "karaoke" is also an inevitable step for building real-world application systems.

## 6. REFERENCES

[1] J. Sundberg, *The Science of the Singing*. Northern Illinois University Press, 1987.

[2] J. Sundberg, "Singing and timbre," *Music room acoustics*, vol. 17, pp. 57–81, 1977.

[3] C. E. Seashore, "A musical ornament, the vibrato," in *Proc. Psychology of Music*. McGraw-Hill Book Company, 1938, pp. 33–52.

[4] J. Large and S. Iwata, "Aerodynamic study of vibrato and voluntary "straight tone" pairs in singing," *J. Acoust. Soc. Am.*, vol. 49, no. 1A, p. 137, 1971.

[5] H. B. Rothman and A. A. Arroyo, "Acoustic variability in vibrato and its perceptual significance," *J. Voice*, vol. 1, no. 2, pp. 123–141, 1987.

[6] D. Myers and J. Michel, "Vibrato and pitch transitions," *J. Voice*, vol. 1, no. 2, pp. 157–161, 1987.

[7] J. Hakes, T. Shipp, and E. T. Doherty, "Acoustic characteristics of vocal oscillations: Vibrato, exaggerated vibrato, trill, and trillo," *J. Voice*, vol. 1, no. 4, pp. 326–331, 1988.

[8] C. D'Alessandro and M. Castellengo, "The pitch of short-duration vibrato tones," *J. Acoust. Soc. Am.*, vol. 95, no. 3, pp. 1617–1630, 1994.

[9] D. Gerhard, "Pitch track target deviation in natural singing," in *Proc. ISMIR*, 2005, pp. 514–519.

[10] K. Kojima, M. Yanagida, and I. Nakayama, "Variability of vibrato -a comparative study between japanese traditional singing and bel canto-," in *Proc. Speech Prosody*, 2004, pp. 151–154.

[11] I. Nakayama, "Comparative studies on vocal expressions in japanese traditional and western classical-style singing, using a common verse," in *Proc. ICA*, 2004, pp. 1295–1296.

[12] G. de Krom and G. Bloothooft, "Timing and accuracy of fundamental frequency changes in singing," in *Proc. ICPhS*, 1995, pp. 206–209.

[13] M. Akagi and H. Kitakaze, "Perception of synthesized singing voices with fine fluctuations in their fundamental frequency contours," in *Proc. ICSLP*, 2000, pp. 458–461.

[14] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Speech-To-Singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices," in *Proc. WASPAA*, 2007, pp. 215–218.

[15] T. L. Nwe and H. Li, "Exploring vibrato-motivated acoustic features for singer identification," *IEEE Transactions on Audio, Speech, and Language processing*, pp. 519–530, 2007.

[16] T. Nakano, M. Goto, and Y. Hiraga, "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," in *Proc. Interspeech*, 2006, pp. 1706–1709.

[17] H. Mori, W. Odagiri, and H. Kasuya, "F0 dynamics in singing: Evidence from the data of a baritone singer," *IEICE Trans. Inf. and Syst.*, vol. E87-D, no. 5, pp. 1086–1092, 2004.

[18] N. Minematsu, B. Matsuoka, and K. Hirose, "Prosodic modeling of nagauta singing and its evaluation," in *Proc. SpeechProsody*, 2004, pp. 487–490.

[19] L. Reqnier and G. Peeters, "Singing voice detection in music tracks using direct voice vibrato," in *Proc. ICCASP*, 2009, pp. 1658–1688.

[20] Y. Ohishi, M. Goto, K. Itou, and K. Takeda., "A stochastic representation of the dynamics of sung melody," in *Proc. ISMIR*, 2007, pp. 371–372.

[21] M. Goto, K. Itou, and S. Hayamizu, "A real-time filled pause detection system for spontaneous speech recognition," in *Proc. Eurospeech*, 1999, pp. 227–230.