Long-Term F_0 Modeling for Text-Independent Speaker Recognition

Tomi Kinnunen and Rosa González Hautamäki

Department of Computer Science
Speech and Image Processing Unit
University of Joensuu, Finland
{tkinnu,rgonza}@cs.joensuu.fi

Abstract

Long-term F_0 modeling for text-independent speaker recognition is considered using both parametric and nonparametric approaches. In the parametric case, mean, variance, skewness, and kurtosis are computed and the parameter vectors are compared using weighted Euclidean distance. In the nonparametric case, F_0 distribution is represented by a histogram, and Kullback-Leibler distance is used in addition to the Euclidean distance. F_0 models are combined with a spectral classifier based on MFCC coefficients, and the results on a subset of the NIST 1999 corpus indicate that F_0 provides useful additional information, especially for improving verification accuracy in noisy and mismatched training/matching conditions.

1. Introduction

Speech fundamental frequency, or F_0, is defined as the rate of vibration of the vocal folds during voiced speech sounds [9]. The F_0 value depends both on the mass and size of the vocal folds, as well as their stiffness and stress [21]. Females and children have smaller vocal folds than adult males, and consequently, their F_0 is higher in general. One of the main advantages of F_0 compared to other speech parameters is that it can be reliably extracted from noisy speech [6, 13, 10]. Even if the first partial is missing (e.g. telephone bandwidth), the F_0 can be still detected according to the upper harmonic structure [8].

Temporal variation of F_0, or the *pitch contour*, plays an important role in signaling emotion and attitude. Pitch contour is also known to be a useful speaker cue, and it was applied in text-dependent speaker recognition already in the 1970's [3]. Local dynamics and piecewise linear modeling of pitch contour has been studied in [13, 20, 1]. One problem with the pitch movements is that they are easy to imitate according to some studies [2].

The long-term F_0 statistics, on the other hand, are more closely related to the physical properties of the larynx. Especially mean F_0 has been studied [14, 16, 6], and it is used in forensics [18, 4]. In [6], variance, skew and kurtosis were also used. Sometimes F_0 is replaced by $\log(F_0)$ [19, 7, 20], which has some perceptual motivations. In [19] it is shown that $\log F_0$ follows normal distribution under some general assumptions.

In this study, we compare and combine parametric and nonparametric approaches to long-term F_0 modeling, and combine the F_0 models with spectral classifier based on MFCC coefficients. We study the complementariness of F_0 with the spectral features on clean and noisy conditions in both the identification and verification tasks.

2. Long-Term F_0 Modeling

The input for F_0 modeling is the observation sequence $X = \{F_0^{(1)}, F_0^{(2)}, \ldots, F_0^{(T)}\}$, extracted from T voiced frames. We consider both *parametric* and *nonparametric* models, as well as their combination.

2.1. Parametric Model

For the parametric model, we compute the mean (μ), variance (σ^2), skewness (γ_1) and kurtosis (γ_2). They are estimated as follows [18]:

$$\begin{aligned}
\mu &= \frac{1}{T}\sum_{t=1}^{T} F_0^{(t)} \\
\sigma^2 &= \frac{1}{T-1}\sum_{t=1}^{T}\left(F_0^{(t)} - \mu\right)^2 \\
\gamma_1 &= \Big(\frac{1}{T-1}\sum_{t=1}^{T}(F_0^{(t)} - \mu)^3\Big)\Big/\sigma^3 \\
\gamma_2 &= \Big(\frac{1}{T-1}\sum_{t=1}^{T}(F_0^{(t)} - \mu)^4\Big)\Big/\sigma^4.
\end{aligned}$$

Mean value characterizes the average F_0, and reflects the "most typical" F_0 value [14, 16, 6]. Variance measures dispersion around the mean value, and is related to the range of F_0 values. Skewness measures the asymmetry of the F_0 distribution (to left or right), and kurtosis is a measure of "peakedness" of a distribution. Kurtosis also measures *nongaussianity* of the distribution.

In the training phase, the parameter vector $(\mu, \sigma^2, \gamma_1, \gamma_2)$ is stored. In the matching phase, the same parameter vector is computed from the test sample, and weighted Euclidean distance is used as the dissimilarity measure of the two distributions.

2.2. Nonparametric Model

For the nonparametric model, we estimate the density function of F_0 using a histogram. In the training phase, we store the raw F_0 values. In the matching phase, data from the reference and test samples is pooled together, and histogram bin positions are determined by dividing the data range into M bins. Using the M common bins, the reference and test densities $p_r(m), p_t(m), m = 1, 2, \ldots, M$ are obtained. In the matching phase, both the Euclidean distance and the *Kullback-Leibler* distance are used for measuring the dissimilarity of the histograms. The Kullback-Leibler distance, or *relative entropy*, is computed as [5]

$$\mathrm{KL}(p_r, p_t) = \sum_{m=1}^{M} p_r(m)\big[\log p_r(m) - \log p_t(m)\big].$$

KL distance is nonnegative and zero for identical distribution, but asymmetric since $\mathrm{KL}(p_r, p_t) \neq \mathrm{KL}(p_t, p_r)$ in general. We use the symmetric form $\frac{1}{2}[\mathrm{KL}(p_r, p_t) + \mathrm{KL}(p_t, p_r)]$ as the dissimilarity measure.

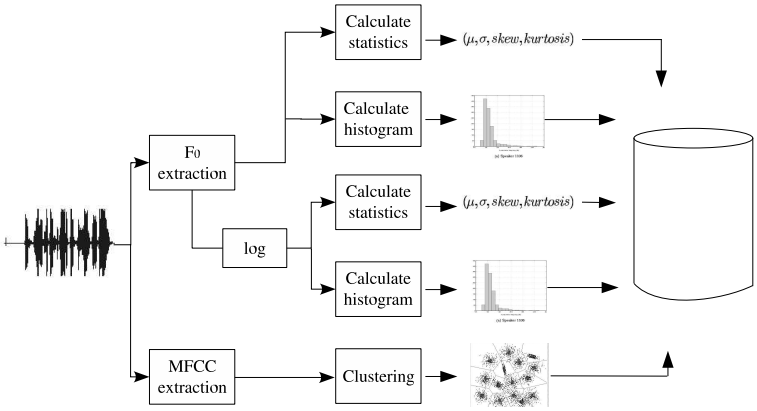

Figure 1: Creating multiparametric speaker profile.

2.3. Combining Different Models

Although computed from the same data, parametric and nonparametric models are based on different assumptions, and they can provide complementary information. We consider the following six models:

F_0-param	:	F_0, parametric
F_0-Eucl	:	F_0, nonparametric, Eucl. dist.
F_0-KL	:	F_0, nonparametric, KL dist.
$\log F_0$-param	:	$\log F_0$, parametric
$\log F_0$-Eucl	:	$\log F_0$, nonparametric, Eucl. dist.
$\log F_0$-KL	:	$\log F_0$, nonparametric, KL dist.

The models are considered both individually and in combination. The combination is formed by weighted sum of normalized distances [11]. In addition to F_0, we add a spectral classifier based on vector quantization modeling (VQ) of mel-frequency cepstral coefficients (MFCC) [12]. The creation of the speaker profile is summarized in Fig. 1.

3. Experiments

3.1. Speech Material

For the experiments, male subset from the one speaker detection task training files of the NIST 1999 SRE corpus [15] is used. The dataset consists of 230 speakers, and the material is conversational speech collected over the telephone network. The original speech samples are sampled at 8 kHz with 8-bit μ-law resolution. For each speaker, there are "a" and "b" files, which are from two different recording sessions. The "a" files are used as the training samples and "b" files as the test samples. The length of both samples is about 1 minute. The autocorrelation algorithm of the Praat software [17] is used for F_0 detection.

The 230 speakers are divided into two disjoint sets, the first 50 speakers (in alphabetical order on the filenames) are used as the tuning set for optimizing the number of histogram bins and the combination weights of the classifier pool. Closed-set identification error rate is used as the objective function in tuning. The optimized parameters are then fixed and the performance is evaluated on the remaining 180 speakers. The number of histograms bins were set as follows: F_0-Eucl = 27; F_0-KL = 17; $\log F_0$-Eucl = 15; $\log F_0$-KL = 65. These were kept fixed for all conditions, but the classifier weights were optimized for noisy and mismatched conditions separately.

3.2. Results

First, we study the individual F_0 models and their combination for both clean and noisy conditions. The "clean" condition refers to the original NIST samples. The noisy condition refers to additive factory noise[1] of 10 dB signal-to-noise ratio. Closed set identification error rates and equal error rates (EER) are reported in Table 1, and the receiver operating curve (ROC) is shown in Fig. 2.

In general, the error rates are high, which is expected. In the identification task, the differences between the methods are relatively small. However, in the verification task, the nonparametric methods clearly outperform the parametric approach. The Kullback-Leibler distance for $\log F_0$ gives the best results in most cases, and it is also robust against noise. A possible reason for the success of $\log F_0$-KL is that the number of free parameters is larger than for other models, and the distributions can be better discriminated. Fusing all six F_0 models reduces error rates slightly in the verification task, and the combined six F_0 models are used for the rest of the experiments.

Table 1: Error rates (%) for F_0 models.

	Identif. error		EER	
Model	Clean	Noise	Clean	Noise
F_0-param	92.8	93.3	38.3	41.1
F_0-Eucl	95.0	96.1	28.3	28.7
F_0-KL	93.9	93.3	26.7	28.4
$\log F_0$-param	93.9	95.0	41.1	42.5
$\log F_0$-Eucl	93.9	96.7	28.3	28.7
$\log F_0$-KL	89.4	89.4	27.5	27.8
Fusing all six	88.9	90.6	27.3	27.2

Next, we study the potential of F_0 as a complementary feature to the MFCC features. Table 2 shows the correlation coefficients between the F_0 distances and the MFCC distances. The correlations are close to zero in all cases, suggesting that the features can provide truly complementary information.

[1]Noise data was obtained from `http://spib.rice.edu/spib/select_noise.html`

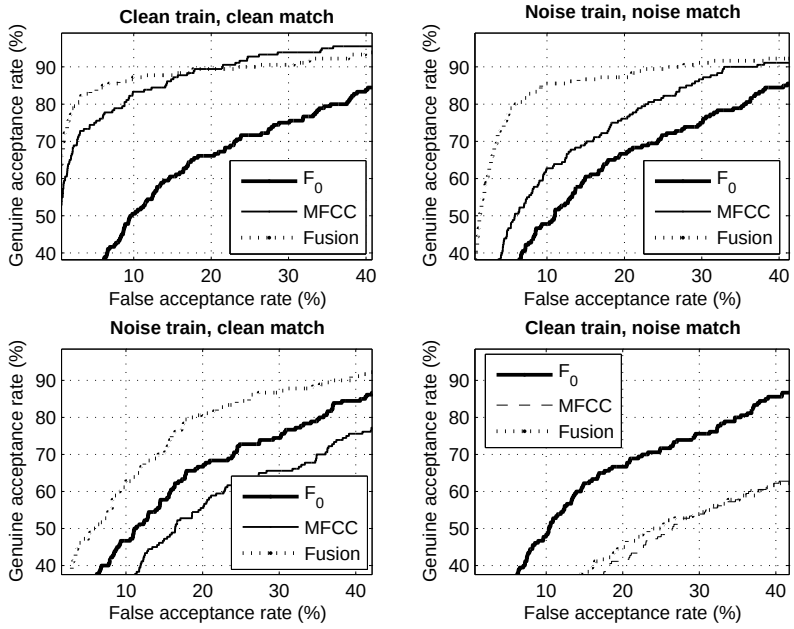

Figure 3: ROC curves of F_0, MFCC and their combination in different conditions.

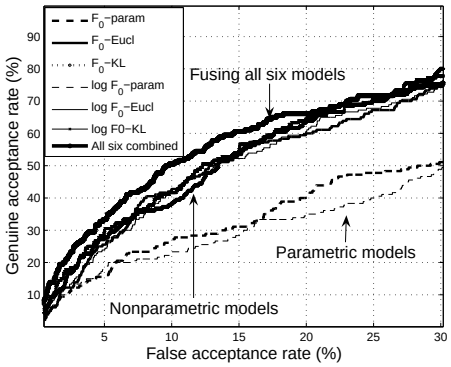

Figure 2: ROC curves for different F_0 models.

Table 2: Correlations between F_0 and MFCC distances.

	MFCC
F_0-param	-0.124
F_0-Eucl	-0.030
F_0-KL	-0.016
log F_0-param	-0.118
log F_0-Eucl	-0.034
log F_0-KL	-0.013

The recognition results are summarized in Table 3 for both matched and mismatched conditions, and the ROC plots are shown in Fig. 3. It can be observed that the accuracy of the MFCC classifier degrades dramatically in noisy and mismatched cases, whereas F_0 is practically unaffected. In the mismatch cases, F_0 verification outperforms MFCC verification.

Fusing MFCC and F_0 yields surprisingly high relative improvements, given that the accuracy of F_0 is rather poor. This is probably due to the uncorrelatedness of the features. For instance, in the noise-noise case, the genuine acceptance rate is increased from 50 % to 80 % at FAR = 5 % by adding F_0. In the noise-clean case, the identification error rate is reduced from 77.2 % to 67.8 %. In the clean-noise case, the fusion is not successful in verification, although it improves identification. A possible reason for this is that the combination weights were optimized for the identification task, and are not guaranteed to improve verification accuracy.

Table 3: Error rates for F_0, MFCC and their fusion.

	Train	Match	F_0	MFCC	Fusion
Identification error rate					
Matched	clean	clean	88.9	23.9	22.8
	noise	noise	90.6	34.4	28.3
Mismatched	noise	clean	90.0	77.2	67.8
	clean	noise	88.9	91.7	86.1
Verification error rate (EER)					
Matched	clean	clean	27.3	14.3	12.2
	noise	noise	27.2	21.7	13.9
Mismatched	noise	clean	27.2	32.9	19.4
	clean	noise	26.4	38.9	38.8

3.3. Samples Misclassified by MFCC but Correctly by F_0

There are a few speakers that were misclassified by MFCC but correctly by F_0, and the speaker labels of these cases are listed in Table 4. For comparative purposes, we show $\log F_0$ distributions along with the long-term average spectra (LTAS) for two speakers in Fig. 4. LTAS shows mismatches in both intensities and spectral shapes, but the $\log F_0$ distributions are very close to each other. This observation supports the independence of spectral and F_0 features.

4. Conclusions

Parametric and nonparametric models for long-term F_0 distribution were studied as an additional cue in text-independent speaker recognition. The results on telephone quality corpus of male speakers indicated that the nonparametric approach is more accurate in general. We recommend to use the nonpara-

Table 4: Speakers misclassified by MFCC but correctly by F_0.

	Train	Match	Speaker labels
Matched	clean	clean	4242, 4402, 4633 4785, 4949, 4996
	noise	noise	4270, 4402, 4487 4543
Mismatched	noise	clean	4270, 4391, 4402 4996, 4999
	clean	noise	4237, 4241, 4270 4391, 4402, 4487 4531, 4535, 4610 4621, 4841, 4914 4949, 4999

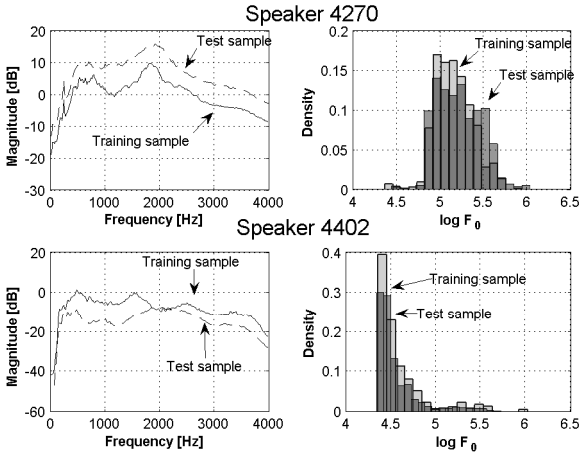

Figure 4: Long term average spectra (left) and F_0 histograms (right) for two speakers.

metric approach as it does not make assumptions about the F_0 distribution. Finding the correct histogram size is one problem in the nonparametric approach, and automatic determination of the histogram size would be an interesting future direction.

It was found out that F_0 is robust against additive noise and mismatch, whereas the accuracy of MFCC features is easily affected by these factors. When used alone, F_0 is rather poor feature (EER 27-42 %). However, when combined with spectral feature, it leads to significant reductions in error rates, especially in the verification task.

5. References

[1] A.G. Adami, R. Mihaescu, D.A. Reynolds, and J.J. Godfrey. Modeling prosodic dynamics for speaker recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, pages 788–791, Hong Kong, 2003.

[2] G. Ashour and I. Gath. Characterization of speech during imitation. In *Proc. 6th European Conf. on Speech Communication and Technology (Eurospeech 1999)*, pages 1187–1190, Budapest, Hungary, 1999.

[3] B. Atal. Automatic speaker recognition based on pitch contours. *Journal of the Acoustic Society of America*, 52(6):1687–1697, 1972.

[4] A. Butcher. Forensic phonetics: Issues in speaker identification evidence. In *Inaugural Int. Conf. of the Institute of Forensic Studies*, Prato, Italy, 2002.

[5] J. Campbell. Speaker recognition: a tutorial. *Proc. of the IEEE*, 85(9):1437–1462, 1997.

[6] M.J. Carey, E.S. Parris, H. Lloyd-Thomas, and S. Bennett. Robust prosodic features for speaker identification. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP 1996)*, pages 1800–1803, Philadelphia, Pennsylvania, USA, 1996.

[7] Y. Cheng and H.C. Leung. Speaker verification using fundamental frequency. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP 1998)*, pages Paper 0228 on the CD–ROM, 1996.

[8] D. Gerhard. Pitch extraction and fundamental frequency: History and current techniques. Technical Report TR-CS 2003-06, University of Regina, Canada, November 2003.

[9] W. Hess. *Pitch Determination of Speech Signals*. Springer-Verlag, Heidelberg, 1983.

[10] K. Iwano, T. Asami, and S. Furui. Noise-robust speaker verification using f_0 features. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP 2004)*, volume 2, pages 1417–1420, 2004.

[11] T. Kinnunen, V. Hautamäki, and P. Fränti. On the fusion of dissimilarity- based classifiers for speaker identification. In *Proc. 8th European Conf. on Speech Communication and Technology (Eurospeech 2003)*, pages 2641–2644, Geneva, Switzerland, 2003.

[12] T. Kinnunen, E. Karpov, and P. Fränti. Real-time speaker identification and verification. *IEEE Trans. on Speech and Audio Processing*, (Accepted for publication).

[13] Y.J. Kyung and H.-S. Lee. Text independent speaker recognition using microprosody. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP 1998)*, Sydney, Australia, 1998.

[14] J.D. Markel, B.T. Oshika, and jr. A.H. Gray. Long-term feature averaging for speaker recognition. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 25(4):330–337, 1977.

[15] A. Martin and M. Przybocki. The NIST 1999 speaker recognition evaluation - an overview. *Digital Signal Processing*, 10:1–18, 2000.

[16] F. Nolan. *The Phonetic Bases of Speaker Recognition*. Cambridge University Press, Cambridge, 1983.

[17] Praat: doing phonetics by computer. www page, July 2005. `http://www.praat.org/`.

[18] P. Rose. *Forensic Speaker Identification*. Taylor & Francis, London, 2002.

[19] M.K. Sönmez, L. Heck, M. Weintraub, and E. Shriberg. A lognormal tied mixture model of pitch for prosody-based speaker recognition. In *Proc. 5th European Conf. on Speech Communication and Technology (Eurospeech 1997)*, pages 1391–1394, Rhodes, Greece, 1997.

[20] M.K. Sönmez, E. Shriberg, L. Heck, and M. Weintraub. Modeling dynamic prosodic variation for speaker verification. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP 1998)*, page Paper 0920, 1998.

[21] I. Titze. *Principles of Voice Production*. Prentice Hall, Englewood Cliffs, NJ, 1994.