Speaker identification based on the statistical analysis of f0

Pavel Labutin, Sergey Koval, Andrey Raev
Speech Technology Center, Saint Petersburg, Russia
koval@speechpro.com

Analysis of pitch, or rather its measurable substrate fundamental frequency (f0), is a generally accepted component of speaker identification decision within both automatic and non automatic speaker recognition. Because the investigation of speech must be comprehensive and because pitch reflects important properties of the human voice, its analysis also should be an obligatory part of the whole investigation. Classic texts on forensic speaker identification use f0 both in automatic and non-automatic speaker identification [5,8,12,13,15,16]. However, the majority of suggestions are for general, relatively simple statistical parameters for pitch curves such as average, range and variation of f0 values [1-4, 6, 9-11]. Our approach is based on many years of experience with f0 data for forensic applications. An algorithm for f0 detection and statistical analysis has been developed and implemented into standard STC software SIS v.6.1.2 or later. F0 is calculated by our two-pass-method using summation of multiple harmonics in the spectral field. This method is adapted for speech signal of very low quality and has also yielded good results for analogue and digital telephone channels.

For forensic work the analyst should select speech produced with a similar emotional intention. Pitch detection quality is controlled by superposition of the calculated pitch curve on the cepstrogamm (i.e. signal periodicity degree function [7]) of the processed signal by means of the SIS software. The analyst can make manual adjustments to the automatically calculated curve in order to remove errors. For speech signals of standard telephone quality mistakes in pitch detection are infrequent. Values of pitch are transformed to a logarithmic scale, and then statistical pitch features are calculated.

The typical set of the statistical parameters measured are: average, maximum, minimum, maximum -3% *, minimum +1%, median, percent of areas with raising f0*, f0 logarithm variation*, f0 logarithm distribution asymmetry*, f0 logarithm distribution excess, average velocity of f0 change, f0 logarithm variation derivative, f0 logarithm derivative distribution asymmetry, f0 logarithm derivative distribution excess, average velocity of f0 rise* and average velocity of f0 fall*. The asterisk indicates the more heavily weighted statistical features.

A speaker identification algorithm was developed and trained using the STC corpus [14]. This RUSTEN speech corpus includes analogue telephone speech. The corpus contains dialogues of 126 speakers (67 women and 59 men) in 5 sessions using 5 different phone analog lines, plus about 1000 files of digital phone channel conversations of 130 speakers. The deviation of every statistical parameter was calculated for every file pair from the corpus. On the basis of these results the distributions of the deviations for pairs "same-different" and "same–same" were built; and functions *false acceptance* (FA), *false rejection* (FR) and EER (equal error rate) were calculated for every statistical parameter. In order to make general identification decisions a common metric was constructed as a weighted sum of separate statistical parameters. The weights were selected to minimize ERR for a given speech database. For the common weighted metric and the common identification metric the deviation distribution for pairs "same-same" and "same-different", FR and FA curves, and ERR were calculated.

Tables 1 and 2 show the results of the investigation of the method using the STC speech data base. In both cases the test data base includes about 1600 speech files of 256 speakers using both analog and digital channels).

Table 1. ERR for speaker identification task using multiple f0 statistics.

Tonal speech duration		10 sec template	20sec template	40sec template	80sec template
10 sec test	All men	17.7 25.2			

	women	26.6			
20 sec test	All men women	16.7 23.7 24.9	15.2 21.7 22.6		
40 sec test	All men women	16.1 23.0 23.8	14.4 20.6 21.1	13.2 19.1 19.0	
80 sec test	All men women	15.6 22.1 23.1	13.6 19.5 19.8	12.3 17.8 17.5	10.9 16.2 15.0

In an analog only telephone net (including public telephones on noisy streets) an 80 x 80sec speaker recognition test with male speakers only produced an EER of about 19%.

Table 2. Equal error rate (ERR) for speaker identification task using only average f0.

Tonal speech duration		10 sec template	20sec template	40sec template	80sec template
10 sec test	Men	32.0			
20 sec test	Men	31.1	30.1		
40 sec test	Men	30.5	30.1		
80 sec test	All Men	30.1	28.8	27.9	17.4 27.5

Conclusion: A method of forensic speaker identification is described which relies upon the statistical analysis of f0. The reliability of the method is tested on a large sample of real speech material consisting of telephone conversations. The method is readily implemented by software, SIS v.6.1.2 or later.

References

[1] Boss D. 'The problem of F0 and real-life speaker identification: a case study', 1996, *Forensic Ling.,* 3(1), pp 155–169.
[2] Braun A. 'Fundamental frequency – how speaker-specific is it?', 1995, *BEIPHOL 64: Studies in Forensic Phon.* 9–23.
[3] Farahani, F., Georgiou P.G., Narayanan, S.S. Speaker identification using supra-segmental pitch pattern dynamics. IEEE International Conference on Acoustics, Speech, and Signal. Processing, 2004. 1, 17-21 May 2004, pp 89-92.
[4] Jiang, M. 'Fundamental frequency vector for a speaker identification system', *Forensic Linguistics, 1996,* 3(1),95–106.
[5]Hollien, H. *The Acoustics of Crime. The New Science of Forensic Phonetics*, 1990, New York: Plenum.
[6] *Jessen M., Koster O, Gfroerer S.* Influence of vocal effort on average and variability of fundamental frequency. 2005,v.12(2), pp 174-213.
[7] Koval S., Bekasova V., Khitrov M., Raev, A. "Pitch detection reliability assessment for forensic applications", In Proc. EUROSPEECH-97, 1997, pp. 489-492
[8] Künzel, H. J. *Sprechererkennung: Grundzüge Forensischer Sprachverarbeitung*, 1987. Heidelberg: Kriminalistik Verlag.
[9] Künzel, H. Effects of voice disguise on speaking fundamental frequency. The International Journal of Speech Language and the Law.2000, v.7 (2),pp 150-179.
[10] Ladd D. R., Terken, J. 'Modelling intra- and inter-speaker pitch range variation', *Proceedings of the 13th International Congress of Phonetic Sciences*, Stockholm, 1995, pp 386–389.
[11] Nolan F. Intonation in speaker identification: an experiment on pitch alignment features. *Forensic Linguistics.* 2002, 9(1), pp 1-21.
[12] Nolan F. Speaker identification evidence: its forms, limitations, and roles. Proceedings of the conference 'Law and Language: Prospect and Retrospect', December 12-15 2001, Levi (Finnish Lapland).
[13] Rose, P. J. *Forensic Speaker Identification*, 2002, London: Taylor &Francis.
[14] RUSTEN: Russian Switched Telephone Network speech database (STC), 2003. S0050, *ELDA - Evaluations and Language resources Distribution Agency*, http://www.elda.fr/catalogue/en/speech/S0050.html.
[15] Galiashina E.I. Forensic phonoscopic examination. 2001, Moscow: Triada (in Russian).
[16] Guidance for forensic speech records examination. D106.2. Ed. Koval S. St. Petersburg: STC. 2000 (in Russian).