SPEAKER IDENTIFICATION USING SELECTIVE COMPARISON OF PITCH CONTOUR PARAMETERS

Natalia Smirnova[1], *Alexey Starshinov*[1], *Ilya Oparin*[1], *Tatiana Goloshchapova*[2]

[1]Speech Technology Center, St. Petersburg, RUSSIA
[2]Federal Service of Drug Control of the Russian Federation, Moscow, RUSSIA
{nsmirnova,starshinov,ilya}@speechpro.com

ABSTRACT

A method of selective pitch data comparison for speaker identification is presented. Pitch parameters of rising and falling nuclear monosyllables and filled hesitation pauses are evaluated for their discriminating ability using Fisher's ratio and equal error rate measures obtained on a 10-male 3-session database of neutral-mode spontaneous statements in Tajik. "Physical" pitch parameters, the best of which provided 75-78% speaker discrimination accuracy in isolation, proved more effective than linguistically conditioned ones. Using all pitch parameters in combination produced identification accuracy of 87%. The ultimate aim of the present work is to develop an algorithm for automated expert assessment of overall pitch pattern similarity in speech samples for its further application within a multifold speaker identification system.

Keywords: speaker identification and verification, pitch contour parameters, forensic phonetics.

1. INTRODUCTION AND BACKGROUND

Considering the variability of pitch parameters and their strong dependence on the stylistic and emotional content of speech, pitch-based forensic speaker identification is generally used to a very limited degree.

The use of pitch parameters in comparing stylistically similar speech samples is complicated by the lack of data on within-speaker stability and between-speaker variability of the pitch parameters used. The F0 minimum of a final low tone is probably the best studied parameter in this respect. Its stability has been verified in a number of studies, some of them involving cases of pitch range mismatch (see [1:64-67] for a detailed review). The mean and standard deviation of F0 are two other commonly used prosodic identification parameters [2]. Of linguistic parameters used in intonation analysis, the alignment of pitch targets proved useful in discriminating between regional varieties of Swedish, English and German [3, 4, 5]. However, little success was achieved when using this feature in between-speaker discrimination [6]. Recently another alignment parameter "mid-fall" has been introduced in a study involving falling-rising pitch contours [6] as potentially useful for between-speaker discrimination.

The present work continues research along this line and provides statistical measures for the stability of pitch values obtained in nuclear rising and falling monosyllables and filled hesitation pauses in samples of stylistically neutral spontaneously produced statements.

2. METHOD

2.1. Theoretical framework

The approach to pitch contour representation adopted for the described method includes elements of the British and Dutch intonation analysis schools [7, 8] and pitch parameters commonly employed in speech intonation research.

The basic analysis unit is a tone-group pitch contour – a complex structure consisting of tonal elements of maximally three types: the pre-head, the head and the nucleus. The nucleus is the only obligatory element in the tone-group. Each of the elements is described by a specific set of parameters depending on its type and sub-type. All parameters are rendered by numerical values and roughly correspond to conventionally used pitch descriptive categories. Thus, rising and falling nuclear monosyllables, which are the subject of the present paper, are described with the following parameters:

- *F0 start value* – the initial F0 value of the nuclear syllable in Hz;
- *F0 end value* – the final F0 value of the nuclear syllable in Hz;
- *F0 maximum* – the maximum F0 value within the nuclear syllable in Hz;
- *F0 minimum* – the minimum F0 value within the nuclear syllable in Hz;

- *F0 mean value* – the arithmetic mean of F0 values within the nuclear syllable in Hz;
- *Timing of F0 maximum* – time point of F0 maximum measured in % of the total nuclear syllable duration;
- *Timing of F0 minimum* – time point of F0 minimum measured in % of the total nuclear syllable duration;
- *Timing of F0 mid-value* – time point of F0 mid-value (lying halfway between F0min and F0max) measured in % of the total nuclear syllable duration. This parameter (as a mid-fall value) was first used by F. Nolan in [6] for the analysis of between-speaker differences in a falling-rising tone.
- *Range* – the interval between F0min and F0max values; measured both in Hz and in semitones;
- *Rate of pitch change* – mean rate of pitch change (rise or fall) within the nuclear syllable measured in Hz/msec. This parameter corresponds to the descriptive category "steep-gradual".

Experimental testing of the suggested methodology on speech data is described in the following section.

2.2. Experiment

2.2.1. Speech material and subjects

In order to test the discriminative ability of pitch parameters from the above given list, an experiment was carried out using speech samples produced by 10 male native speakers of Tajik in 3 recording sessions. All subjects spoke educated Tajik with slight traces of regional accent.

A brief characterization of speakers, including 3-session long-term pitch means, is given in Table 1.

Table 1: The speakers.

Speaker	Age	Dialect	Mean F0, Hz
SP1	25	Dushanbe (West)	110…114
SP2	21	Dushanbe (West)	134…139
SP3	20	Dushanbe (West)	113…119
SP4	35	Dushanbe (West)	114…115
SP5	18	Dushanbe (West)	107…110
SP6	35	South	130…140
SP7	51	South	106…112
SP8	35	North	110…123
SP9	38	South-East	127…138
SP10	35	South-East	114…115

The recorded speech material included a naturally produced autobiography and a short narrative about the speaker's place of origin.

The sound was recorded via a Sennheiser PC 160 headset microphone in digital format directly into a PC at a 22050 Hz sampling rate, 16 bit. The time interval between successive sessions was one week or longer.

The data was processed using a specially designed pitch analysis module performing automatic calculation of all relevant pitch parameters within a selected pitch contour fragment.

Speech fragments for analysis were selected on the basis of their comparability by a non-native Tajik-speaking trained phonetician capable of producing "verbatim" transcriptions of the analysed speech fragments using a Tajik dictionary. The most important factors taken into account were emotion, vocal effort and utterance mode comparability (e.g. rising monosyllables in non-final clauses can be very different in their characteristics from question rises). Another factor considered was segmental context comparability. Thus, to reduce the effect of segmental context on pitch values, only monosyllables having the "voiced consonant + vowel" and "voiced consonant + vowel + voiced consonant" structure were selected.

Besides rises and falls, it was decided to use filled hesitation pauses (HP) as a separate unit of analysis, because they appeared quite often in the analyzed speech material.

In each speaker session from 5 to 20 comparable samples of each unit type were selected. The proportion of rises and falls differed from speaker to speaker, mainly due to individual preferences for either rising or falling termination of final and non-final clauses. For filled pauses only 3 parameters were analyzed – F0max, F0min and F0mean. Since Speaker 6 produced no filled pauses in any of his sessions, statistical analysis relating to this unit could be performed for 9 speakers only. Tables with each parameter statistics, including their means and variance, were obtained automatically.

2.2.2. Data analysis and results

The discriminative potential of statistical parameters is commonly evaluated by the so-called F-ratio (Fisher's ratio) calculated as a ratio of between-speaker variance of parameter value means to mean within-speaker variance of the same parameter values using the following formula:

$$F = \frac{\sigma_1^2}{\sigma_2^2}, \qquad (1)$$

where σ_1^2 is between-speaker dispersion of parameter mean values, while σ_2^2 is mean within-speaker dispersion of the same parameter values.

In order to see how consistent the speakers were in pitch contours within one session, F-ratios were first obtained for within-speaker dispersion of all parameter measures in the speech sample, while between-speaker variance was calculated for the means only. The results for the two types of nuclei (fall and rise) and for hesitation pauses (HP) are presented in Table 2.

Table 2: F-ratio of between-speaker means variance to within-speaker within-session variance.

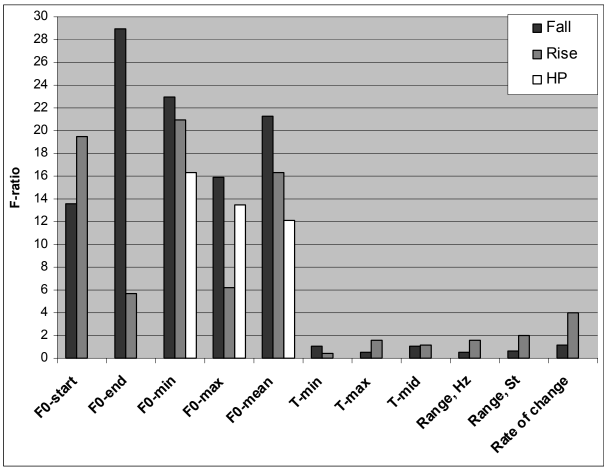
Parameter	Fall	Rise	HP
F0-start	2,6	3,5	-
F0-end	4,5	1,4	-
F0-min	4,4	3,7	8,8
F0-max	2,9	1,4	5,5
F0-mean	3,8	2,2	8,3
T-min	0,3	0,1	-
T-max	0,1	0,2	-
T-mid	0,2	0,5	-
Range, Hz	0,2	0,5	-
Rate of change	0,3	0,5	-

Thus, potentially useful as between-speaker discriminators seem to be only the so-called physical F0 parameters, while for all linguistic pitch parameters intra-speaker variability exceeds inter-speaker variability. Of the physical F0 parameters the most promising proved to be those of filled hesitation pauses – mainly due to very low within-speaker variance.

Meanwhile, closer analysis of speaker strategies in pitch unit realization revealed a strong tendency for all speakers to generally prefer a particular way of nuclear pitch realization in similar contexts, so that parameter means showed high consistency across sessions. So at the next stage it was decided to use speaker means obtained for 3 sessions (rather than all parameter measurements) for calculating within-speaker variance. The resulting F-ratio values are shown in Fig. 1.

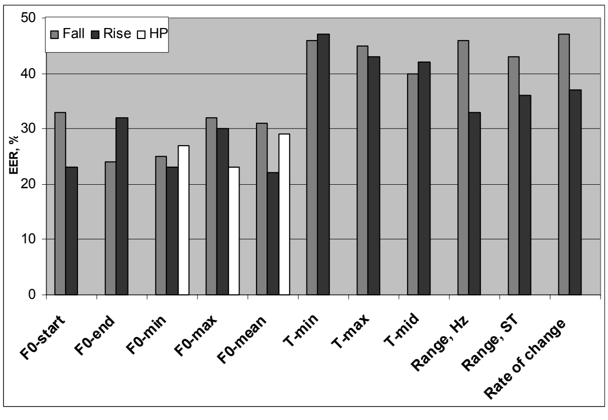
The same tendency for physical F0 parameters to have more potential as between-speaker discriminators is observed, except that the values increased in most cases. Thus, the end F0 value of a fall has the highest F-ratio of 29. Of non-physical parameters, the only tangible improvement was achieved for the rate of pitch change of a rise – from 0,5 to 4. Pitch range of a rise measured in semitones produced slightly better results than when measured in Hz (2 vs. 1,6). However, for most linguistic parameters F-ratio remained low - either slightly over or even below 1 (timing of F0 maximum of a fall, timing of F0 minimum of a rise and the range of a fall).

Figure 1: F-ratio of between-speaker means variance to within-speaker variance of 3-session means.



The actual effectiveness of the described pitch parameters for between-speaker discrimination was tested using EER (equal error rate) obtained on the same speech database at parameter thresholds giving equal false acceptance and false rejection error rates. The number of within-speaker comparisons per parameter was 30 for rises and falls (10 speakers*3 sessions) and 27 for filled hesitation pauses (9 speakers*3 sessions). The number of between-speaker comparisons was 405 per parameter for rises and falls and 324 for filled hesitation pauses. The EER values were derived as means of the closest FA and FR values. The results for parameters in isolation are presented in Fig. 2.

Figure 2: EER obtained on a 10-speaker 3-session speech database for parameters of rising and falling nuclear monosyllables and filled hesitation pauses.



Quite in conformity with F-ratio values, error rates are lower for physical F0 parameters than for linguistic ones. Thus, most of the physical parameters have the reliability over 70%, while the effectiveness of linguistic parameters is in most cases below 60%. Ineffective for speaker discrimination within the analyzed speaker population were timing of F0 minimum (both of falls and rises), pitch range (Hz) and rate of pitch change of falls.

Six parameters (of the total of 25) produced EER from 22% to 25%. Contrary to expectations, rises performed in most cases better than falls. The fusion of only rise parameters (excluding timing of F0 minimum) produced an EER of 15%. The addition of fall and filled pause parameters slightly improved the result still further reducing the overall EER to13%.

3. CONCLUSIONS AND FUTURE WORK

The preliminary results reported in this paper demonstrate that despite their extreme variability, parameter values of locally determined pitch events still provide useful information on some speaker-specific aspects of pitch contour realization, at least in stylistically comparable speech samples.

The so-called physical F0 parameters proved to be better in between-speaker discrimination, most of them producing EER between 22% and 30%. The physical parameters of rises, falls and hesitation pauses were not in most cases equally correlated across speakers, which partly explains the fact that the combination of parameters produced an overall drop of EER to 13%.

The rather discouraging fact that linguistic pitch parameters, in particular timing, performed so poorly, can to some extent be explained by the monosyllabic structure of the analyzed nuclei allowing too little space for feature variation. If this is indeed the case, the situation could improve on nuclear structures with at least one post-nuclear syllable. However, in this case care should be taken to accurately discriminate between two categorically distinct pitch accent types – early vs. late rises and falls. Meanwhile, closer analysis of data shows that even though non-physical F0 parameters generally allow for considerable intra-speaker variance, they are still effective in a limited number of cases when speakers are relatively consistent in the realized values while showing clearly expressed differences from each other or deviations from typical realizations. Thus, in the analyzed data two speakers (SP4 and SP10) could be reliably differentiated by their pitch rise ranges (6-7 ST vs. 3,5-3,7 ST). Likewise, SP3 and SP10 could be regularly differentiated by their mid-fall values (24-43% vs. 57-71%). The optimal use of such parameters will thus consist in capturing the most evident between-speaker differences, as well as deviations from "standards" accepted in the relevant language community.

Having in view the prospect of method application in forensic tasks, it should be emphasized that the findings reported here apply only to neutral-mode spontaneous utterances and cannot be extended to samples involving emotional, communicative or vocal effort mismatch.

At the present stage effective method application requires a specially trained forensic expert to perform the task of selecting speech data, in particular the control of vocal effort, emotional and communicative nuance, syllabic structure and segmental makeup of speech fragments. Currently developed automatic data analysis procedures are designed to check the usability of data and to optimize data sets for comparison.

Further method development includes, among others, increasing the number of pitch units, structure types, speakers and language types for comparison, introducing additional parameters, in particular durational features and ratios and derivatives of some parameters. An important issue to be researched is the stability of the used parameters in conditions of style (i.e. read vs. spontaneous) and vocal effort mismatch.

4. REFERENCES

[1] Ladd, D. R. 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

[2] Künzel, H. J. 1987. *Sprechererkennung: Grundzüge Forensischer Sprachverarbeitung*. Heidelberg: Kriminalistik Verlag.

[3] Bruce, G., Frid, J., Thelander, I. 2004. Swedish Accent Navigation. *International Symposium on Tonal Aspects of Languages: Emphasis on Tone Languages,* Beijing, 9-12.

[4] Nolan, F., Farrar, K. 1999. Timing of F0 peaks and peak lag. *Proc. 14th ICPhS* San Francisco 2, 961–4.

[5] Peters, J. 1999. The timing of nuclear high accents in German dialects. *Proc. 14th ICPhS* San Francisco 3, 1877-1880.

[6] Nolan, F. 2002. Intonation in speaker identification: an experiment on pitch alignment features. *Forensic Linguistics* 9(1), 1-21.

[7] O'Connor, J., Arnold, G. 1973. *Intonation of colloquial English*. London: Longman.

[8] 't Hart, J., Collier, R., Cohen, A. 1990. A Perceptual Study of Intonation: An experimental-phonetic approach to speech melody. Cambridge: Cambridge University Press.