Travaux du 19ème CIL | 19th ICL papers
Congrès International des Linguistes, Genève 20-27 Juillet 2013
International Congress of Linguists, Geneva 20-27 July 2013

19th International
Congress of Linguists
July 21-27 2013
Geneva - Switzerland

Anatole FIODOROV

Minsk State Linguistics University, Belarus
lotana@anitex.by

Professor D. Crystal's Acoustic Profile

poster presentation in session: 4 Phonology and Morphology (Marc van Oostendorp)

Published and distributed by: Département de Linguistique de l'Université de Genève, Rue de Candolle 2, CH-1205 Genève, Switzerland
Editor: Département de Linguistique de l'Université de Genève, Switzerland
ISBN:978-2-8399-1580-9

PROFESSOR D. CRYSTAL'S ACOUSTIC PROFILE

Anatole Fiodorov

Minsk State Linguistics University, Belarus
lotana@anitex.by

Abstract: In the present paper an effort has been made to describe the speaker's style by statistical methods. The syntactic structure of the text along with the formant tracking, F0 analysis and timing of the prepared speech contribute significantly to the idiosyncrasy of the speaker. The data obtained lead to believe that speaker D.C. controls negligible changes in relative syllabic durations to lend an emphasis to a word in the syntagm. The reliability of different statistics is discussed along with syntactic peculiarities of the speaker's idiosyncrasy.

1. INTRODUCTION

The researchers will never cease to look for useful parameters that may help to reliably describe the speaker both statistically and acoustically.

According to Professor Crystal, 'in speaker verification, a sample of speech is acoustically analysed to check a claimed identity against a stored reference sample stored in the computer' [2]. The systems for speaker's voice specification are of keen interest worldwide. It is caused by various practical applications among which are: verification of access to the channels of communication and computer databases, bank accounts, means of transport and its control, oral speech instructions and military devices [10].

The voice channel offers a reasonable edge over fingerprints, credit cards, signatures, and it may be used over a distance, through radio and telephone lines. Voice identification/verification is frequently used as forensic evidence in court and analysis of talk registration onboard air- and spaceship [6].

This problem involves the specification of distinctive features or parametric description of a speech sample, and creation of the speaker's profile for verification tasks [7]. There were numerous investigations seeking the most robust technique and parameters to accomplish these ends [12]. The present study addresses this problem not only in terms of indicative F0 contours and spectrograms but linguistic message as well.

2. MATERIAL and METHOD

A series of talks prepared by Professor David Crystal (speaker D.C.) for a BBC programme [5] were chosen for the present study. The recordings were analysed with the help of PRAAT [2] and PASW Statistics-18 [8] computer programs. The material was segmented manually into syllables and measured by using audio feedback, oscillograms and spectrograms. If a stretch of sound did not lend itself to precise syllable segmentation, it was taken as one stretch [9]. The accuracy of duration measurements was within 5 – 15 ms per word and the fundamental frequency was averaged over vocalic segments instrumentally.

Four sentences, cut out from the radio talk "Clueless" [5], were subjected to further research of prosodic parameters – timing and pitch configuration:

1) I'm **totally** clueless when it comes to *grammar*.
2) I'm **not** totally clueless when it comes to ***grammar***.
3) I'm **totally** clueless when it comes to **cooking**.
4) I'm **totally** *clueless*!

The speaker's average F0 reflects such individual characteristics, among others, as vocal cord size, emotional and health state, gender, etc. Long-term parameters are important because they are considered as the useful indicators of individual voice quality [7]. In this study the rate of speech was associated with the average duration of a syllable as a unit of speech production. The overall speed of pronunciation was represented by the average duration of a syllable in the passage, i.e. the overall time of sounding text divided by the number of syllables in the 'ideal' transcription.

The syntagm is a minimal meaningful rhythmic unit in speech production. The boundaries of syntagms are usually signalled by a slowdown of speech rate or pre-pausal lengthening. When marking syntagm boundaries we relied mostly on auditory analysis after iterative listening to the passage in question.

Apart from the arithmetic average F0, we considered the harmonic mean and such descriptive statistics as median and mode, coefficient of variation *Cv* and standard deviation of the averaged duration or frequency.

We also undertook a study of superimposed formant tracks obtained from another three speakers, engaged for the same BBC program, who pronounced the word 'language' in their talks. The formants were compared visually but were not analyzed quantitatively since they did not produce any noticeable distinctions.

3. RESULTS and DISCUSSION

The comparison of relative syllable *duration structures* with the patterns of relative *F0 movement* in the same utterances has shown no significant correlation between them. However, when the temporal and melodic patterns for similar utterances were considered separately, we obtained marked rank correlation up to 0.70 per cent.

If the dynamics of relative duration signals the boundaries of syntagms by pre-pausal lengthening, the pitch contours are hardly predictable: lengthy portions of static modal F0 values give way to stretches with higher or lower F0 values.

As evident from Table 1 and 2, there are significant correlations between the test utterances although they were uttered with emphasis on different words (they are marked by bold and italic fonts). At the same time the data suggest that the similarity in timing patterns does not necessarily imply correlation of F0 patterns.

Table 1: Spearman rank correlations between syllable relative timing patterns of utterances from the talk "Clueless" produced by the speaker D.C. Marked correlations are significant at $p < 0.05$.

Utterance	1	2	3
1. I'm **totally** clueless when it comes to *grammar*.	1.0		
2. I'm **not** *totally* clueless when it comes to *grammar*.	**0.70**	1.0	
3. I'm **totally** clueless when it comes to **cooking**.	0.39	**0.69**	1.0

Table 2: Spearman rank correlations between syllable relative F0 patterns of utterances from the talk "Clueless" produced by the speaker D.C. Marked correlations are significant at $p < 0.05$.

Utterance	1	2	3
1. I'm **totally** clueless when it comes to *grammar*.	1.0		
2. I'm **not** *totally* clueless when it comes to ***grammar***.	0.55	1.0	
3. I'm **totally** clueless when it comes to **cooking**.	**0.70**	**0.65**	1.0

The descriptive statistics in Table 3 suggest that the pitch modulation can be treated as one of the most stable characteristics of the speaker's voice. This points to the fact that the speaking tessitura, a given speaker's range of comfortable speaking pitch, serves as another useful indicator of the speaker's idiosyncrasy.

Table 3 indicates that the speaker D.C. prefers temporal dynamics to pitch contour fluctuations: Cv = 53.3% against 30.8% respectively. At the same time it is safe to say that the average F0 (harmonic rather than arithmetic mean) is likely to be another reliable factor for speaker verification.

Table 3: Descriptive statistics of the speaker D.C.'s idiostyle based on the BBC recordings about 8 min long.

Descriptive statistics				
Syllable, *ms*		Syntagm, *ms*	F0 statistics, *Hz*	
Mean	230.4 ± 122.8	756.6 ± 368	Mean (arithmetic)	98.87 ± 30.5
Median	204.5	726	Mean (harmonic)	93.3± 28.4
Moda	242.0	multiple	Median	89.0
Coef. of var. (Cv)	53.3%	48.6%	Coef. of var. (Cv)	30.8%

Since we usually observe in speech numerous outliers beyond the statistical confidence limit, we have good grounds for believing that the median of syllable durations is far more robust than the arithmetic mean.

The *syntagmatic structure* of an utterance may serve as an integral indicator of the subject's mentality, attitude and idiosyncrasy in general. Although in this study the syntagm length is subject to sizable variation (48.6%), the median can be regarded as a useful tool for speaker verification.

Table 4 shows that the descriptive statistics of a sentence length leave a little chance to believe that the length of an utterance in words can be an indicative feature of idiostyle.

Table 4: Descriptive statistics of sentence length (words)

Mean	Median	Mode	Frequency of Mode	Standard Deviation	Coef. Var.
20.0	17.0	multiple	6	12.3	61.6%

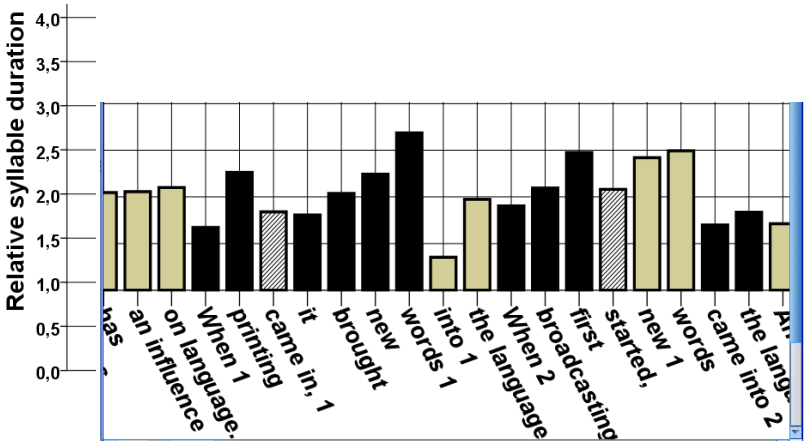
Nevertheless, several *syntactic* and *prosodic* peculiarities may be attributed to the idiostyle of the speaker D.C. Among such characteristics are the following.

(1) Primarily, **parenthetic words** uttered with a marked temporal downtrend: *just for fun; … sang a song about it, is not just because of …*

(2) The pronunciation of certain syntagms in two rhythmic groups is another peculiar feature of the speaker's style worth mentioning:

back | in the seventies and eighties …; *back | in the 1930-ies …*

(3) As illustrated in Fig. 1, there is a pronounced tendency to organize syntagms as meaningful units by slowing down the pace, although the last word in the syntagm can be uttered much quicker. This points to the fact that the syntagm division is not necessarily marked with a prolonged syllable of a word.



*...when printing **came in**;*
*...when broadcasting first **started**;*
*...with every **item**;*
*...especially the last ten years **really**;*
*...on the restaurant **menu**;*
*...you might expect **to see**.*

Fig. 1: A fragment of the relative syllable duration chart for the text "Spam": same colour bars denote syntagms complete with grey bars marking their boundaries.

(4) The same rule applies to **compound verbs** when a preposition/adverb is pronounced rather quickly: *came **along***; *printing came **in***. Yet, this observation does not hold good for the final position in the sentence where the preposition/adverb is being prolonged: *...and it caught **on***.

(5) The author employs **iterations** in the sentence as one of his favourite expedients: *new words* × 2; *the word spam* × 2; *based upon it* × 2; *since* × 2, the last word being usually prolonged before specification of the time to follow.

(6) Similarly the speaker tends to employ prosodic and syntactic **reduplications**, which are characterised by identical pitch contours and timing patterns of corresponding words:

originally tinned meat ⇔ kind of cold meat;

when printing came in ⇔ when broadcasting first started;

verbs based upon it ⇔ adjectives based upon it;

of any kind ⇔ kind of ⇔ other kinds of;

who do the work themselves ⇔ who send all these horrible emails.

These contrasts and parallelisms serve as a means for emphasizing important points. Figure 2 demonstrates an example of pitch contours with a pronounced agreement between the prosodic patterns of similar utterances, complete with the onset / offset values of the corresponding curves.

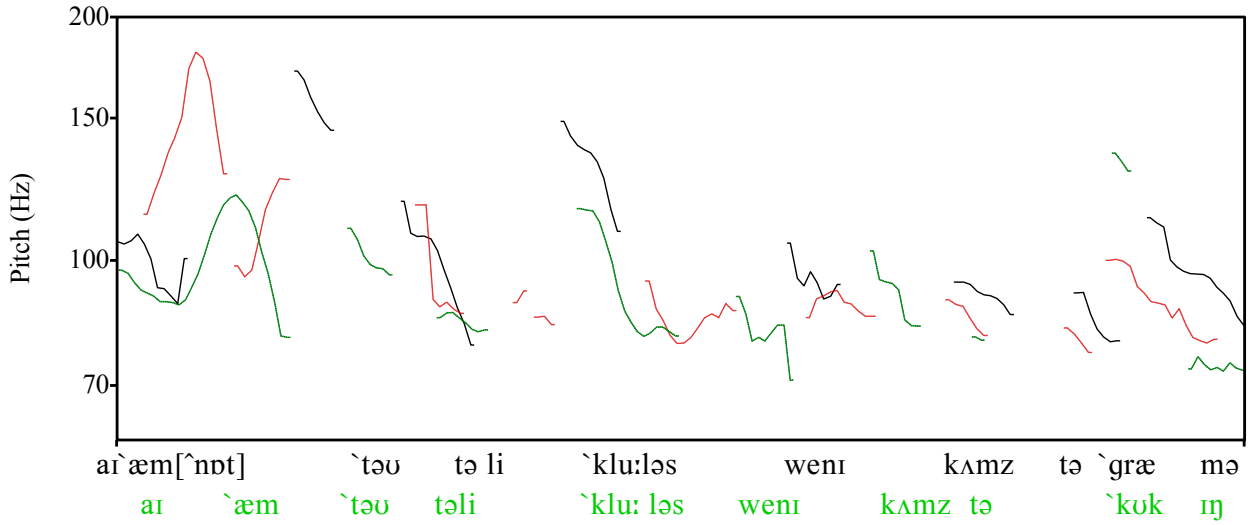


Fig. 2: The ensemble of three superimposed (non-aligned) pitch contours for the test utterances: the solid black curve – No1, the dotted red curve – No2, the dashed green curve – No3 (see the list of sentences above).

For *spectrographic analysis* we have chosen the word ‘whatever’. It is rather common in casual English and, as a rule, it is given an emphasis in the utterance, which makes the formants easier to pinpoint. Figure 3 shows the superimposed in one window 12 formant tracks of the test word cut out from the talk “Whatever”, dealing with the usage of the popular word [5].

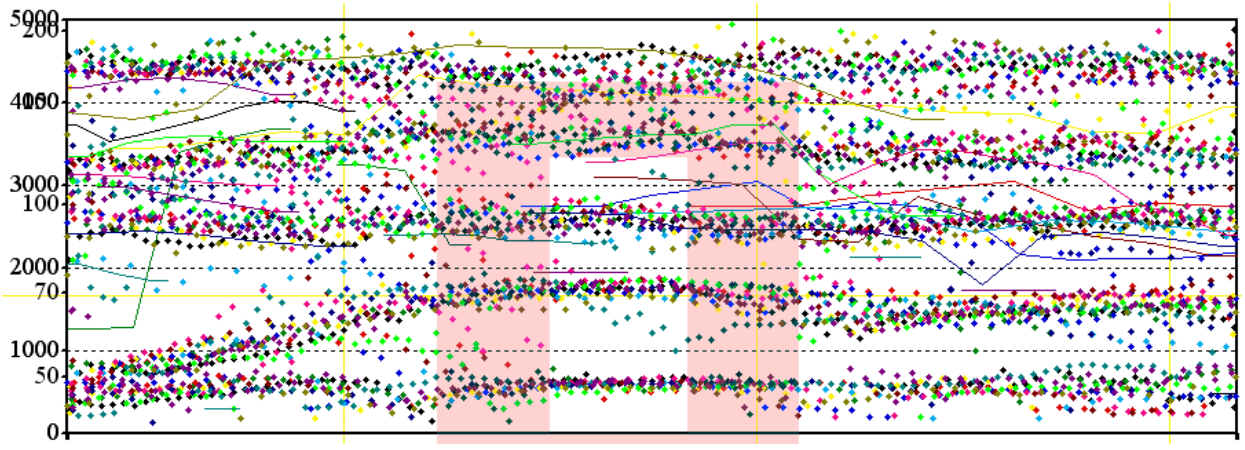


Fig. 3: Five formant tracks on the word ‘whatever’ [wɒ`(t)evə] recorded by speaker D.C. in different sentences (12 occurrences). The stationary portion of [e] is marked by the window. The solid curves show F0 contours. The figures represent the formant frequency (1000 – 5000 Hz) and F0 (50 – 200 Hz).

Figure 3 depicts the narrow range of frequency variations for the stressed vowel [e]: F1 = 617 ± 47.5 Hz; F2 = 1780 ± 91.0 Hz with low coefficients of variation – 7.7% and 5.1% respectively. A few outliers on this stretch are within random mistakes of the readings provided by the program. These data lead us to believe that the calculated values of F1 and F2 can contribute as a useful clue to the speaker verification task.

4. CONCLUSION

The investigation reported in this paper indicates a number of problems involved in the speaker verification. Some factors that influence the selection of a particular parameter have been described. It appears that the idiosyncratic portrait of a speakers should derive both from the acoustic, syntactic and semantic features. It is difficult to say which variables are most useful, nevertheless different linguistic and acoustic databases may contribute to the solution when treated in a package.

The data reported here suggest that the median of fundamental frequency and relative syllable duration are quite informative for the verification of the speaker. The fine temporal structure of an utterance seems to be another leading factor for the speaker verification task. Among other useful features one should mention the degree of pitch contour fluctuation and the length of syntagm in terms of median values and coefficients of variation.

The investigation confirmed the relevance of statistical rank correlation analysis applied to the comparison of relative timing and pitch structures. Along with other parameters the spectrographic templates of certain test words should be included into the speaker's verification database.

The described features need to be verified statistically on a larger experimental basis involving a greater sample of subjects. In conclusion, we have to agree that "a significant part of between-speaker variability exhibits regularities which are linguistic, rather than solely acoustic, in nature" [1].

5. REFERENCES

[1] Barry, W. J., Hoequist, C. E. and Nolan, F. J. 1989. An approach to the problem of regional accent in automatic speech recognition', *Computer Speech and Language*, vol. 3, p. 355 – 366.

[2] Boersma, P, and Weenink, D. 2014. Praat: doing phonetics by computer. Version 5.3.82, retrieved 26 May 2013 from http://www.praat.org/

[3] Crystal, D. 2008. *A Dictionary of Linguistics and Phonetics* / David Crystal. – 6th ed. – Wiley-Blackwell.

[4] Fiodorov, A. 1987. Native or Alien: Verification of Foreign Accent in the Speech of Russian Learners of English. *11th ICPhS* Tallinn, vol. 5, 142-145.

[5] Keep your English up to date 1-6, retrieved 27 May 2012 from http://www.bbc.co.uk/worldservice/learningenglish/radio/specials/

[6] McDougall K., Nolan F. 2007. Discrimination of Speakers Using the Formant Dynamics of /u:/ in British English. *Proc*. *16th ICPhS* Saarbrucken, 1825-1828.

[7] Nolan, F. 1997. Speaker recognition and forensic phonetics. Hardcastle, W.J., Laver, J. (eds.), *The Handbook of Phonetic Sciences*. Oxford: Blackwell, p. 744-767.

[8] PASW Statistics 18.0, retrieved 27 May 2012 from http://pasw-statistics.software.informer.com/18.0/.

[9] Roach, P. 1991. *English Phonetics and Phonology: A Practical Coarse* / Peter Roach. – 2nd ed. – Cambridge University Press.

[10] Rose P. 2002. *Forensic Speaker Identification* / Philip Rose. London and New York: Taylor & Francis.

[11] Smirnova N., Starshinov A. et al 2007. Speaker Identification Using Selective Comparison of Pitch Contour Parameters. *Proc*. *16th ICPhS* Saarbrucken, 1801-1804.