https://doi.org/10.3176/lu.1998.3.27

ANNA-LIISA LEHESSAARI (Helsinki), *RAIMO TOIVONEN* (Tampere)

PANORAMA — PROSPECT — PROFILE: ON THE COMPUTER-AIDED DESCRIPTION OF PROSODIC QUALITY OF SPEECH

Technical advances have, as a rule, become important milestones in the history of phonetic research. Describing the prosodic quality of speech requires illustrative methods, quantitative data, and statistical analyses. From a phonetic viewpoint, not only objective acoustic measurements (perceived phenomena) are needed but also subjective semantic knowledge of prosodic structures of communicational significance (labeled forms). However, if no device is available for speech recording and analysis, prosodic phenomena can hardly be understood — this is, unfortunately, the most common situation for example when immigrants meet difficulties in learning to speak the new language. Because of the multidimensional nature of prosody, a set of complementary methods, rather than one specific method, is needed. In this paper, we discuss a three-level approch to the computer-aided prosodic analysis of recorded speech samples:

level 1) p a n o r a m a, or an overview of the speech sample studied,

level 2) p r o s p e c t — collecting the quantitative and qualitative data,

level 3) p r o f i l e — indicating the observed differences between speech samples.

Our presentation is based on work done using the Intelligent Speech Analyser™ (ISA), a software tool developed and designed by one of the writers, R. Toivonen (Toivonen 1986; for updated details, see Intelligent Speech Analyser™ homepage: www.sci.fi/~pitchsys).

Level 1

P a n o r a m a presents an overview of the speech sample, from the very beginning to the end. The amplitude envelope, accompanied with a loudness curve, might be the most beneficial display (Figure 1). This kind of overview is very helpful when working, for example, with samples of conversational speech.

Since time resolution is controllable, no detail is lost in the panorama. Any speech sequence of interest can be temporally stretched out for further investigation, segmentation, and acoustic analysis. All the analyses have their own display windows. Contents of all the displays can be listened to. Usually, more accurate observations and a better understanding of the material are attained, and time and speculation saved, if several analysis windows are opened instead of a single one.

The term *segment* is used here in a special meaning. Once a segment is identified by determining its beginning and its end in one display window, it is updated to all the display windows. Segments can be named and re-named. If necessary, segment boundaries can be corrected or removed. The sound file can be saved with

segment information. Different levels of segmentation are available; for example, one level for syllables and another for utterances.

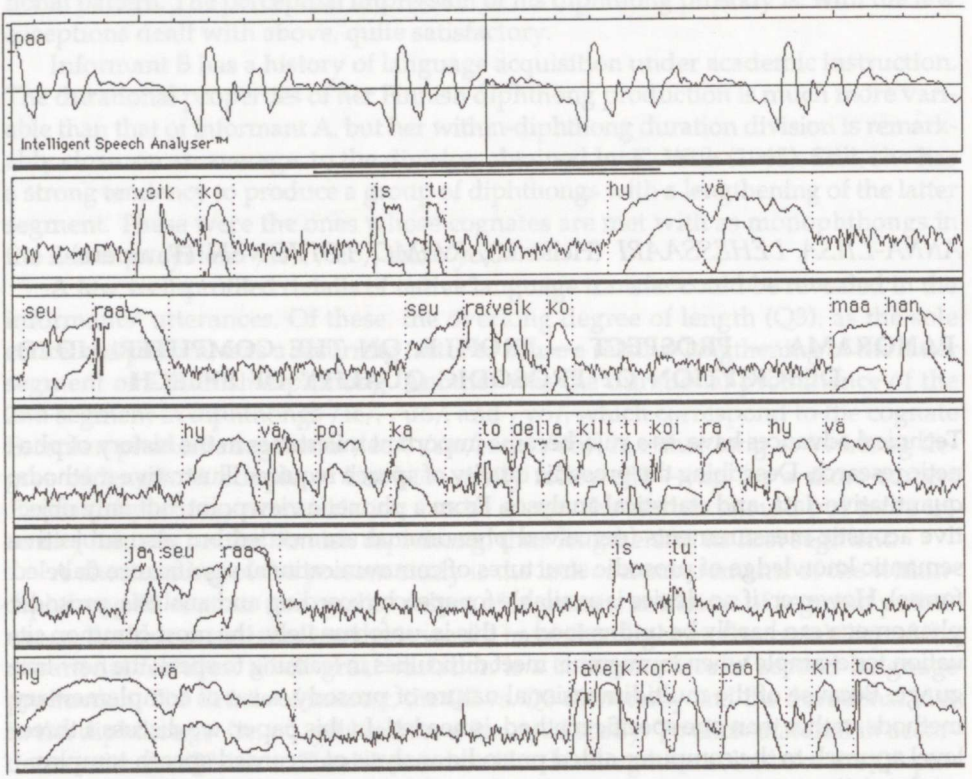

Figure 1. **Panorama: Obedience training with Veikko, a dachshund. Upper window: signal display; lower window: speech amplitude envelope display with loudness curve. For sample information, see table 1. Transcripted:** *Veikko*! *Istu. Hyvä. Seuraa. Seuraa, Veikko*! *Maahan. Hyvä poika. Todella kiltti koira, hyvä*! *Ja seuraa*! *Istu. Hyvä. Ja Veikko on vapaa, kiitos*! 'Veikko! Sit. Good. Follow. Follow, Veikko! Go down. Good boy. Very nice dog, good! And follow! Sit. Good. And Veikko is free, thank you!'

Level 2

P r o s p e c t is the formula for notation, storage, and prosessing which we use for acoustic, auditory, semantic, and other information related to a speech sample (see Table 1). Quantitative and qualitative information is included about the whole speech sample as well as about segments of different levels, such as syllables and utterances.

Combined with statistical functions, the prospect makes the kind of tool that is much needed when working with the massive amounts of data involved in the study of prosodic parameters. For studying variation in a parameter, histogram displays and descriptive statistics are available. Functions for the study of correlations between parameters and among acoustic and perceptual forms are under construction.

Plus and minus tags are used for tag-guided analysis. Segments with plus tags are included in the analysis whereas segments with minus tags are excluded. For example, minus tags can be used to exclude speech sequences disturbed by noise. Or, we can calculate a histogram of the fundamental frequency distribution for all syllables, for the population of stressed syllables, and for the population of unstressed syllables.

Table 1

Prospect chart (an example)

Conversation: Obedience training 24.4.98, outdoors.
Speaker: A.-L. Lehessaari **With:** Veikko the dachsie.
Total: 43 syllables / 12 utterances / duration 29 800 ms.
In this chart: UTTERANCES 1—5.

Syllable level:																				
	veik	*ko*	*	*is*	*tu*	*	*hy*	*vä*	*	*seu*	*raa*	*	*seu*	*raa*	*veik*	*ko*
stress	+			+				+						+		
sp. rate							S	S							F	F
dur.	429	208		352	131		613	703		450	491		372	179	366	166
F0 max	306	177			188		230	185		220	248		253	302	192	160
Utterance level:																
dur.	673		896	483		1045	1316		1023	941		1522	1083			
end	falling			falling			falling			non-falling			non-falling			

(chart continues →)

* = pause; stress = syllable stress, "+" tag; sp. rate = perceived speech rate, S = slow, F = fast; dur. = duration (ms); F0 max = fundamental frequency maximum (Hz); end = utterance-final pitch type.

Level 3

In a study of dysarthria, C. Ludlow and C. Bassich (1983 : 138) used a Z score to "determine the degree of impairment relative to normal performance":

$$Z = \frac{\text{Normal Subjects' Mean} - \text{Patients' Value}}{\text{Normal Subjects' Standard Deviation}}$$

This formula makes it possible to project onto the same figure parameters measured on different scales. The idea of a multiparametric profile has been applied for the prosodic comparison of speech samples, for example, by A.-L. Lehessaari (1996; see Figure 2).

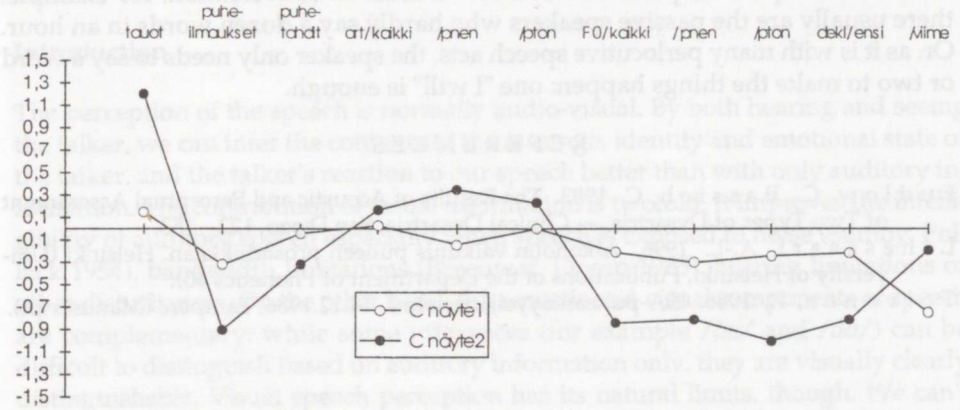

Figure 2. **(Adapted from Lehessaari 1996 : 180). Z scores for 11 speech parameters: slight intoxication (unfilled symbols, speech sample of 223 syllables) and medium intoxication (filled symbols, speech sample of 224 syllables) compared to sober speech (speech sample of 223 syllables). Speaker C, Finnish-speaking female, 25 years. Parameters:** *tauot* **= mean duration, all pauses;** *puheilmaukset* **= mean duration, all utterances;** *puhetahdit* **= mean length in syllables, all speech measures;** *art/kaikki, /pnen, /pton* **= mean duration, all syllables /stressed syllables /unstressed syllables;** *F0/kaikki, /pnen, /pton* **= mean fundamental frequency maximum, all syllables /stressed syllables /unstressed syllables;** *dekl/ensi, /viime* **= 0.5 · mean fundamental frequency maximum, first stressed /last stressed syllable in a declination unit.**

However, there are several open questions of quite fundamental nature. First of all, aren't we trying to make visual those measured physical phenomena which are the psychoacoustically relevant ones? What is the most representative set of parameters for a profile? How to select the best scale for a parameter? These problems are also related to the panorama and to the profile.

Statistical tests of significance in their general form, as used in physics or sociology, are hardly suitable for the study of every phonetic issue. Such tests take into account not only the difference between the averages obtained for the speech samples to be compared but also the magnitude of standard deviation observed. With prosodic phenomena, wide variation is often observed, even expected. Hence, very large speech samples might be needed to obtain significant differences.

Table 2

Mean duration and standard deviation (ms)
(All syllables and all pauses; data adapted from Lehessaari 1996 : 104, 127)
Speaker N, Finnish-speaking female, 33 years, sober and intoxicated

	Sober	Intoxicated	Significance
syllables	165 (69) N = 918	191 (82) N = 888	*** ($p < .001$)
pauses	518 (360) N = 54	807 (489) N = 39	** ($p < .01$)

In this example (Table 2), intoxication is characterised by a clear prolongation of pauses. However, because of the large standard deviation and the small number of pauses, this difference is statistically only "significant" note the relatively smaller but statistically "highly significant" prolongation in syllable duration! The problems often cannot be solved by collecting more material. It might simply be that in a given situation the speakers produce but a few words. In conversation, for example, there usually are the passive speakers who hardly say a dozen words in an hour. Or, as it is with many perlocutive speech acts, the speaker only needs to say a word or two to make the things happen: one "I will" is enough.

REFERENCES

Ludlow, C., Bassich, C. 1983, The Results of Acoustic and Perceptual Assessment of Two Types of Dysartria. — Clinical Dysartria, San Diego, 122—153.

Lehessaari, A.-L. 1996, Alkoholin vaikutus puheen prosodiikkaan, Helsinki (University of Helsinki, Publications of the Department of Phonetics 40).

Toivonen, R. 1986, ISA-puheanalyysijärjestelmä. 14.12.1986, Tampere (Manuscript).