

*SDDD: A new dissimilarity index for the comparison of speech spectra**

Bernard HARMEGNIES

Département de Phonétique et Psychoacoustique, Université de l'Etat à Mons, Avenue du Champ de Mars, Ch. II., B-7000 Mons Belgium

Received 3 June 1988

Abstract: SDDD, an index for the comparison of speech spectra, is introduced in this paper. Its discriminatory ability is established by comparing SDDD with the best classical index on the basis of a speaker-recognition experiment using Long Term Average Spectra.

Key words: Voice, voice quality, dissimilarity index, long term average spectrum.

1. Introduction

Long-Term Average Speech Spectra (LTAS) have been used successfully in several experiments dealing with voice characterization. Various fields have been investigated, e.g., speech pathology [1], deafness [2], language differences [3–5] and speaker-identification [6–12]. In these kinds of research, the first problem authors are faced with is the one of comparing spectra. Several authors [2–5] simply performed a contrastive observation of the spectra. Others [1] associated each LTAS with an index value, so that simple arithmetical reasoning allowed to compare them. Others [6–12] used metrics capable of delivering numerical values indicating the degree of (dis)similarity for the spectra under comparison. For purposes of highly acute voice characterization, the last approach obviously appears to be the best.

In this paper, we introduce a new dissimilarity index, the Standard Deviation of the Differences Distribution (SDDD), which has proved to be quite effective for comparing the shapes of high-dimensional long-term spectra [12]. The efficiency of SDDD will be tested by simulating a speaker-

identification experiment and comparing the respective recognition rates exhibited both by SDDD, and the cross-correlation coefficient, classically regarded as the most practical and powerful index for comparing LTAS [9–12].

2. Indices for the comparison of long term spectra

We consider each LTAS as a k -dimensional vector, with k the total number of frequency channels in the spectrum. Therefore, spectrum \mathbf{S} may be defined as follows:

$$\mathbf{S} = (S_1, \dots, S_i, \dots, S_k) \quad (1)$$

with S_i , the level of the i th frequency component. In this paper as well as in most previous ones [5,8,9,10,12], the S_i values will be expressed in decibels.

2.1. A classical index: the cross-correlation coefficient

The Bravais–Pearson cross-correlation coeffi-

* This research is partially sponsored by the Belgian Fonds National de la Recherche Scientifique under grant number 2.4450.86F.

cient (R) can be used as a similarity index for the comparison of LTAS. It expresses the tendency of the S_i values to covary with the S'_i values and it ranges, in absolute values from 0 (complete independence of the S_i and S'_i variabilities) to 1 (perfect correlation of the S_i and S'_i values). Formally, R can be defined as:

$$R_{ss'} = \frac{1}{k} \frac{\sum_{i=1}^k (S_i - M_s)(S'_i - M_{s'})}{\sigma_s \sigma_{s'}} \quad (2)$$

where M_s and $M_{s'}$ are the means for all S_i and S'_i values, respectively, and σ_s and $\sigma_{s'}$ are the corresponding standard deviations. If the spectra being compared are identical, the correlation is 1; on the contrary, a weak correlation indicates a lack of similarity of the spectral shapes. Among the classical indices that can be found in the LTAS-related literature, the correlation coefficient appears to be one of the more interesting. On the one hand, it has quite high discriminating ability [9]. On the other hand, unlike most other indices (e.g. the Euclidean distance, or the chi-square index), it is insensitive to differences in the overall levels of the compared spectra. Therefore, it does not require any preliminary intensity-normalization. This is a particularly interesting feature, for normalization can be a time-consuming process. Furthermore, choosing one normalization process can be a delicate question since normalization itself may generate important biases [13].

2.2. The SDDD index

The level-differences between the spectra under comparison provide much information about shapes similarities. Particularly, the standard deviation of these differences (SDDD) appears to be of great interest. It is defined as:

$$\text{SDDD}_{ss'} = \sqrt{\frac{1}{k} \sum_{i=1}^k (S_i - S'_i - M_d)^2} \quad (3)$$

where M_d is the average of the $S_i - S'_i$ differences. If the spectral shapes compared are highly similar, the differences values are almost invariant and tend to concentrate around a given central tendency in-

fluenced only by the between-spectra overall level difference. If the shapes are different, large level differences can be found in certain frequency channels and small ones in others. The standard deviation of the differences therefore increases. In Figures 1a and 1b are shown two LTAS drawn from 2 utterances of the same text produced by a single speaker. The statistical distribution of the between-spectra levels differences is shown in Figure 1c. In Figures 1d and 1f, are presented two LTAS drawn from utterances of the same text produced by two different speakers. The corresponding statistical distribution of the differences is pictured in Figure 1f. As shown by Figures 1c and 1f, the dispersion of differences is quite higher in the case of inter-speaker comparisons (here, $\text{SDDD} = 7.154$) than in the one of intra-speaker comparisons (here, $\text{SDDD} = 3.966$). SDDD increases when the similarity of the spectra decreases, and conversely.

Moreover, SDDD exhibits good robustness against changes in the overall levels of LTAS. In order to demonstrate this property, let us consider the comparison of spectra S and Z , both obtained on the basis of two identical speech segments differing by levels only. Provided that the system is linear, each component of spectrum Z can be obtained by adding a given constant, c , to the corresponding component of spectrum S , i.e.

$$Z_i = S_i + c \quad (1 \leq i \leq k). \quad (4)$$

Therefore, $Z_i - S_i = c$ for all i and so does M_d . Accordingly $\text{SDDD}_{sz} = 0$. In fact, transforming a spectrum by formula (4) causes the whole distribution of differences to shift by a distance equal to c ; the shape of the distribution, and consequently its dispersion remain invariant. For those reasons, SDDD is insensitive to changes in the overall levels of the spectra and, therefore, does not require any preliminary normalization.

3. Experimentation

3.1. Experimental setting

The speakers were 10 French-speaking male subjects, between 19 and 21 years old. Each of them

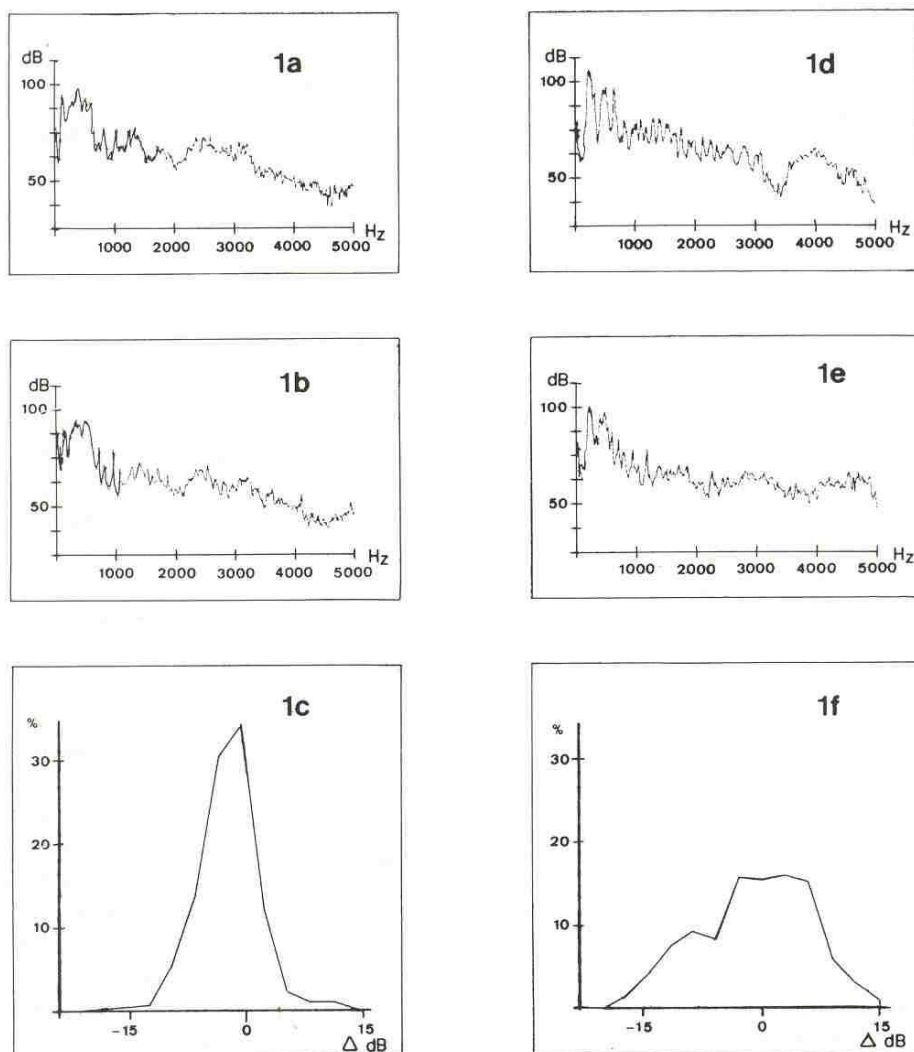


Figure 1. Two pairs of LTAS and the corresponding distributions of interspectral differences: two LTAS drawn from two utterances of the same text by a single speaker (1a, 1b) and their distributions of differences (1c); two LTAS drawn from utterance of the same text produced by two different speakers (1d, 1e) and their distributions of differences (1f).

was asked to utter a phonetically balanced French text ten times in succession. The text was about 18 seconds long. The utterances were recorded on a Nagra IV S recorder, by means of a KM 84 Neumann microphone. The recording sessions took place in a quiet sound-proof room; the subjects were sitting in front of the microphone, placed at a constant 40 cm distance from their lips. The recordings were analyzed by means of a 400-channels FFT analyzer (Brüel Kjaer 2033). Its sampling frequency was set at 12.8 kHz, determining a constant 12.5 Hz

resolution across the whole 0–5 kHz range of analysis. A linear averaging process was used in order to compute one LTAS for each utterance. The so-obtained 100 LTAS were then transmitted from the analyzer to an Apple II computer via a GPIB interface card, for further computations.

3.2. Comparison procedure

Inter- as well as intra-speaker comparisons were performed. For each of the 10 speakers, one com-

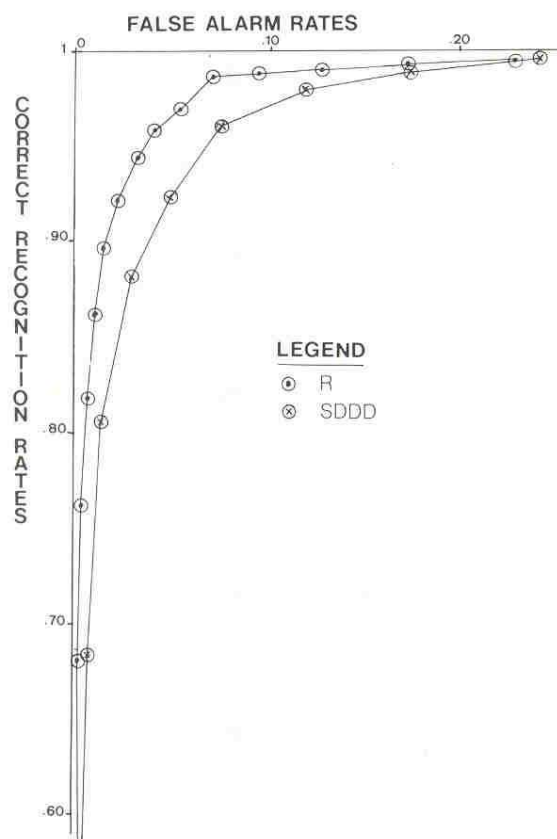


Figure 2. Experimental R.O.C. curves for the cross-correlation and the SDDD indices.

parison was performed for each possible non-redundant pair of his 10 LTAS, i.e., 45 comparisons by subject, thus 450 intra-speaker comparisons for the whole sample. Symmetrically, for each possible non-redundant pair of different speakers (i.e. 45 pairs of speakers), 100 inter-speaker comparisons were achieved, i.e., 4500 inter-speaker comparisons for the whole sample. For each comparison of two spectra, both the indices under investigation were computed.

3.3. Results

For each similarity index, the 450 intra-speaker values and 4500 inter-speaker values computed¹ constituted the inter- and intra-speaker distributions we used to assess the discriminatory ability of the indices.

¹ These results are not mentioned in this paper but will be submitted later for publication.

For each index, a series of values selected across its entire range of variation was successively considered as rejection thresholds for a recognition task. The corresponding false alarm- and correct recognition rates were drawn from the observed distributions: two ROC curves (one by index) were drawn from this statistical processing. They are shown in Figure 2.

4. Discussion

It has been shown [14] that the proportion of the area of the entire ROC space that lies beneath a ROC curve is a distribution-free measure of sensitivity: the greater this surface, the higher the sensitivity. It is evident, from a glance at Figure 2, that the surface under the *R* curve is smaller than the one under the SDDD curve. It is moreover quite particular to see that the curves never cut each other. We can therefore conclude that, at least in the case of this experiment, SDDD exhibits superior between-speakers discriminating ability than the cross-correlation index.

One of the reasons for this superiority can be found in the theoretical concepts underlying the indices under comparison. We have stressed that both were insensitive to changes in the overall spectral levels. In other words, spectra *S* and *Z* will be considered perfectly similar (according to *R* as well as to SDDD), provided that they match equation (4). Nevertheless, it should be emphasized that, as a linear regression-based coefficient, *R* is insensitive to any linear spectral transformation. According to *R*, spectra *S* and *Z* will therefore be considered perfectly similar if they match the more general equation:

$$Z_i = a S_i + c \quad (a \neq 0) \quad (5)$$

On the contrary, SDDD takes a zero value (perfect similarity) only in the case where *a* equals 1, no matter the value of *c*. In this sense, SDDD appears to be sensitive to a class of shape differences for which *R* exhibits a total blindness.

In order to illustrate this property, we have generated 5 artificial 10-channels spectra scaled accord-

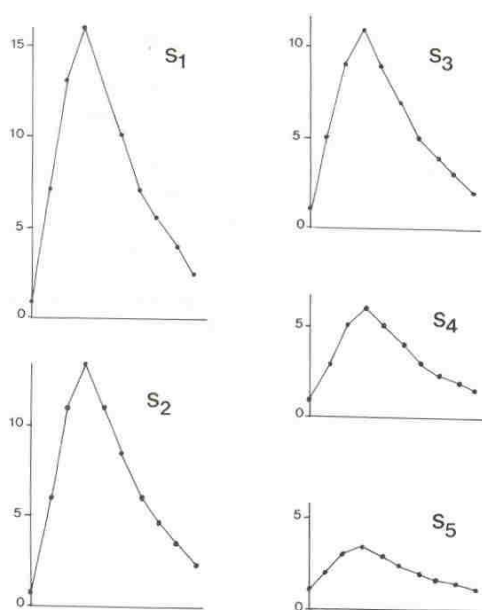


Figure 3. Five artificial 10-channels spectra ($S_1 - S_5$) only differing from one to another by a linear transform of their intensity values.

ing to (5) with $a \neq 1$ and $c = 0$. These spectra are shown in Figure 3. The between-spectra R and SDDD values have been computed and are listed in Table 1. They show that SDDD exhibits values di-

Table 1
Values of R and SDDD (underlined) computed for each pair of the five artificial spectra pictured in Figure 3

	S_1	S_2	S_3	S_4	S_5
S_1	1	1	1	1	1
	<u>0</u>	<u>0.748</u>	<u>1.568</u>	<u>3.137</u>	<u>3.921</u>
S_2		1	1	1	1
		<u>0</u>	<u>0.784</u>	<u>2.353</u>	<u>3.137</u>
S_3			1	1	1
			<u>0</u>	<u>1.568</u>	<u>2.353</u>
S_4				1	1
				<u>0</u>	<u>0.784</u>
S_5					1
					<u>0</u>

rectly linked to the intuitively perceptible differences in spectral shapes, although the computation of R provides in any case a unit-value.

5. Conclusion

The experimental results clearly show the efficiency of SDDD as a dissimilarity index for comparing LTAS. SDDD is obviously more efficient than R . Moreover, we have shown that, like R , SDDD is insensitive to changes in the overall levels of the spectra; this feature may allow to save on processing time and/or storage capacity. Using SDDD in order to compare LTAS thus appears as a good choice both for economy in processing and accuracy in discrimination.

In this paper, SDDD was tested by means of a speaker-recognition experiment involving high-dimensional long-term speech spectra. Presently, other experiments involving spectral comparisons by means of SDDD are being carried out in the phonetics laboratory of the Mons University. We think however that SDDD might be of use for shape comparison in other fields than the one of LTAS or even of speech sciences. This remains to be tested.

Acknowledgments

Thanks are due to Professor Albert Landercy for his constant and friendly help.

References

[1] Frokjaer-Jensen, B. and S. Prytz (1976). Registration of voice quality. *Bruel Kjaer Technical Review* 3, 3-17.
 [2] Formby, C. and R.B. Mosen (1982). Long-term average spectra for normal and hearing-impaired adolescents. *J. Acoust. Soc. Am.* 71 (1), 196-202.
 [3] Tarnoczy, Th. (1958). Détermination du spectre de la parole avec une méthode nouvelle. *Acustica* 8, 392-395.
 [4] Banuls-Terol (1971). Weighted average spectrum of human speech: an approach. *Acts 7th International Congress on Acoustics*, Budapest, 1971, 253-256.
 [5] Harmegnies, B. and A. Landercy (1985). Language features in the long-term average spectrum. *Revue de Phonétique Appliquée* 73-75, 69-80.

- [6] Furui, S., F. Itakura and S. Saito (1972). Talker recognition by longtime averaged speech spectrum. *Electron. Com. in Japan* 55-A (10), 54-61.
- [7] Furui, S., F. Itakura and S. Saito (1975). Personal information in the long-time averaged speech spectrum. *Review of the Electrical Communication Laboratories* 23 (9-10), 1133-41.
- [8] Majewski, W. and H. Hollien (1974). Euclidean distances between long-term speech spectra as a criterion for speaker identification. In: Fant, G., Ed., *Speech Communication, vol. 3*. Almqvist and Wiksell, Stockholm, 303-310.
- [9] Zalewski, J., W. Majewski and H. Hollien (1975). Cross-correlation of long-term speech spectra as a speaker identification technique. *Acustica* 34, 24.
- [10] Hollien, H. and W. Majewski (1977). Speaker identification by long-term spectra under normal and distorted speech conditions. *J. Acoust. Soc. Am.* 62 (4), 975-980.
- [11] Bunge, E. (1977). Herkenning van sprekers door een computer. *Philips Techn. Rev.* 37 (7), 179-192.
- [12] Harmegnies, B., and A. Landercy (1986). Comparison of spectral similarity indices for speaker recognition, *Proc. 12th ICA, vol. 1*, Toronto, 1986, A1-4.
- [13] Everitt, B. (1980). *Cluster Analysis*. Halsted Press, New York.
- [14] Swets, T.A. (1973). The relative operating characteristic in psychology. *Science* 182, 990-1000.