https://doi.org/10.3176/lu.1998.3.06

ANTTI IIVONEN, TUIJA NIEMI-LAITINEN, KIRSI HARINEN (Helsinki)

EVALUATION OF SIMILARITY DEGREE BETWEEN SPEAKERS ON THE BASIS OF SHORT TIME FFT SPECTRA

1. Introduction

We concentrate on the question, to which extent speaker specific features and individual differences can be found in the short time FFT spectra in the cases in which the speakers to be compared sound very similar. Our paper includes three contributions according to three authors.

We have investigated the effect of the analysis options in order to reveal which options yield the best results for comparison. Certain options are kept unchanged. Because telephone speech is the type of speech most often occurring in the forensic applications, we have included also the real recordings via telephone (Kirsi Harinen) and the simulation of the telephone band (300—3400 Hz; all authors) in the comparison. Other spectral bands were used, too.

It can be assumed that quite big amount of the speaker's individual characteristics is included in the single sound spectra. Because the sound spectra are physical correlates to the phonological entities (phonemes), it is understandable that the spectra of the same phoneme must be similar to certain extent in two speakers. The additional individual difference is interesting for the speaker verification, identification, and discrimination. It would be ideal, if the similarity of the tokens of the same linguistic structure remains great within the same speaker, but gets lower, when another speaker produces the same structure. Different speech sound types may indicate more individual variation than the others (Nolan 1983; Paliwal 1983).

2. Programs and measurements applied in the analysis

We have utilized the SoundScope speech analysis program (GW Instruments) and the option short time FFT spectrum (snap; cf. Altosaar, Meister 1995). In the averaging procedure, the effect of the number of the single snaps was investigated (3 or 6 snaps within a short period of time, e.g. 50 ms). We kept in all analyses the following options unchanged: the sampling rate 22050, the band 5 kHz, the window type Hamming, Pre Emphasis (6 dB/octave). The variable options were: filter band (45 Hz which corresponds to 33 ms time window or 300 Hz/5 ms time window), smoothing (none or low). The single snaps were always taken from the middle portion of a sound (except the plosives in which the bursts were analyzed).

The digital representations of the single spectra were exported (copy wave text) into the Spectral Comparison program, created by means of the FutureBasic II programming language. This program calculates the mean of eigen value differences at different resolution points (within the selected frequency band), their standard deviation and the correlation coefficient (Pearson) between two spectra. It is possible to first calculate an average spectrum of several single spectra and then to compare two averaged spectra with each other. For the comparison of two spectra (or two averaged spectra), the program applies the principle of the best fit by averaging all differences at the measurement points (on the frequency scale) and making one of the two spectra more similar with the other one by subtracting the average value from all values of the other spectrum.

3. Factors affecting the properties of single sound spectra (sources of error)

Among other things, the following factors affect the shape of the spectrum of a single speech sound: the temporal location of the measurement point within the speech sound, the surrounding speech sounds (coarticulation), degree of stress, height of the fundamental frequency, the location within a single period (in resonant sounds), voice quality, emotion, random variation, the type of the analysis option, recording circumstances, recording devices, and speech style. Several single spectra can be gathered at intervals of 10 ms or at the same pitch synchronous measurement points (the single snaps from different periods, but always at the same location within a period) and make an averaged spectrum of those in order to avoid casual and minute (unimportant) variations. Hence, a more stable spectral form can be obtained. When single snaps are averaged, the principle of "the best formant shape visually observed" can be manually applied in the gathering (Figure 1). Otherwise undesired extra variation will be obtained. For obtaining a more stable picture, the smoothing option seems to yield good results.

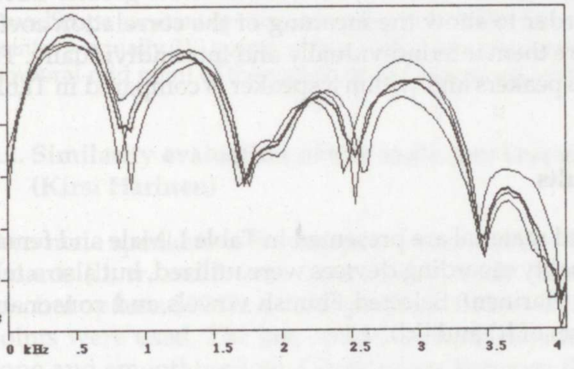

Figure 1a. **Three single spectral (snaps) and the averaged spectrum with "well-shaped" formants and their average spectrum. (A male speaker AI; test vowel [æ] from** *säde* **'ray').**

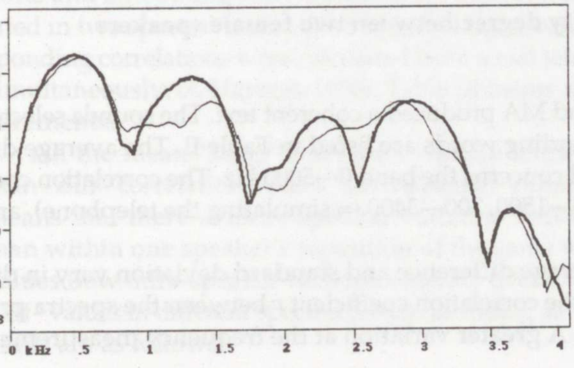

Figure 1b. **A comparison of two averaged spectra: with "well-shaped" (good) and "not well-shaped" (bad) formants.**

Table I

Experimental arrangements (speakers, their linguistic backgrounds, speech material and recording devices) in the three investigations.

experiment.	experiment by					
arrangements	**A. Iivonen**		**K. Harinen**		**T. Niemi-Laitinen**	
code of speaker	HA	MA	MAN	AK	TN	TP
sex, age	female, 19	female, 19	male, 41	male, 34	female, 31	female, 41
linguistic background	Standard Ostro Bothnia	Standard Häme	Standard	North Karelian	Standard South Karelian	Standard South Karelian
material	reading of a coherent text		word list			
phonemes	a selection of consonants and vowels					
recorder	Revox A700 > TEAC W-440-C		Marantz			
microphone	AKG C 451E		AKG acoustics C567 E			
tape	BASF > TDK AD		TDK AD			
telephone	*		DORO 1133		*	
answering machine	*		Panasonic EASA-PHONE		*	

4. Methods used in the comparison

The mean value of the differences at the frequency points and their standard deviation as well as the correlation analysis of the two (averaged) spectra of the two speakers were used in the comparison. Certain number of speech sounds can be analyzed and they can be arranged according to their similarity degree. Certain sounds may show great similarity and certain other sounds show greater difference (cf. Table II below). In order to show the meaning of the correlation coefficients, it is necessary to compare them interindividually and intraindividually. The similarity degree between two speakers and within a speaker is compared in Tables III—IV below.

5. Test arrangements and results

Speakers, recording devices, and material are presented in Table I. Male and female speakers were chosen. High quality recording devices were utilized, but also a telephone answering machine (K. Harinen). Selected Finnish vowels and consonants were analyzed (including bursts of [t] and [k]).

5.1. Evaluation of the similarity degree between two female speakers (Antti Iivonen)

The two female speakers HA and MA produced a coherent text. The sounds selected from the text and the corresponding words are listed in Table II. The average difference of the spectra compared concerns the band 0—5000 Hz. The correlation coefficient *r* of the bands 0—5000, 0—1500, 300—3400 (= simulating the telephone), and 1500—3400 Hz are indicated.

Table II shows that the average difference and standard deviation vary in different speech sounds and that the correlation coefficient *r* between the spectra produced by two speakers varies. A greater variation at the frequency measurement points corresponds to the smaller correlation between the spectra. The same phoneme (cf. the burst of /k/) can give different correlations in different contexts. The best inter-speaker variation was found in [ii]/*liikkuvat*. The closer examination of the corresponding spectra revealed a nasal formant at 1100 Hz in the speaker HA's spectrum. The larger frequency bands showed greater difference between the speakers than the more narrow bands.

Table II

Comparison of the female speakers HA and MA (16 phones within one sentence)

	band 0 — 5000 Hz		**correlation coefficient *r* of different bands**			
phone—word	**av. diff**	**sd**	**0—5000**	**0—1500**	**telephone**	**1500—3400**
[ss]—*parvissa*	2.5	1.5	**.97**	.72	**.93**	.98
[k1]—*kevätkalan*	3.2	2.2	**.90**	.90	**.92**	.97
[a2]—*kevätkalan*	2.8	1.8	**.87**	.78	**.84**	.98
[æ]—*kevätkalan*	2.4	1.6	**.86**	.95	**.74**	.96
[a1]—*kevätkalan*	4.7	2.1	**.81**	.98	**.89**	.97
[mm]—*kalamme*	4.0	2.4	**.81**	.77	**.76**	.92
[v]—*kevätkalan*	3.1	1.4	**.80**	.79	**.74**	.94
[a2]—*kalamme*	4.4	2.4	**.79**	.95	**.84**	.98
[l]—*kevätkalan*	4.9	2.0	**.77**	.70	**.70**	.93
[e]—*kevätkalan*	3.1	1.8	**.74**	.92	**.69**	.96
[t]—*kevätkalan*	4.7	2.3	**.61**	.82	**.55**	.85
[k2]—*kevätkalan*	6.4	3.0	**.58**	.86	**.51**	.88
[k]—*kalamme*	9.0	4.7	**.42**	.86	**.58**	.93
[l]—*kalamme*	8.7	3.8	**.41**	.44	**.18**	.87
[r]—*parvissa*	5.9	3.6	**.34**	.72	**.53**	.93
[ii]—*liikkuvat*	8.7	4.4	**.25**	.72	**.45**	.79
MEANS	**4.9**	**2.6**	**.68**	**.81**	**.68**	**.91**

SoundScope constant options: sampling rate 22.050; filter 45 Hz; TW = 33 ms; resolution 1054 points (= actually 213 points within 5 kHz); low smoothing; PreEmphasis; one snap from the temporal mid point of the speech sound concerned (exept the bursts of [k] and [t]).

5.2. Similarity evaluation of two male speakers and telephone speech (Kirsi Harinen)

Two male speakers to be compared sound alike. Vowels [a, i, ee] as well as consonants [ll, rr, mm] were used in this study (see Table III below). SoundScope spectral options 300 Hz — 1024 points, 300 Hz — 512 points as well as 45 Hz — 1024 points were used. The last option was still devided into two groups: smoothing none and smoothing low. Correlations between the average spectra were calculated in two spectral bands (0—5 kHz and the simulated telephone band) and corresponding correlations were calculated from a real telephone speech sample (recorded simultaneously: cf. Harinen 1996). Table III below shows an example of these measurements.

All the mean "same" correlation values (intra-speaker variation) are greater than "diff" correlation values (inter-speaker variation) in all spectral bands. This means that there is more spectral variation between speakers (same utterance) than within one speaker's repetition of the same word. The highest mean "same" values (low intra-speaker variation, mean *r* over 0.90 = bolded) compared to mean "diff" values in different spectral bands (including all the SoundScope analysis options used) are as follows:

1. 45 Hz — 1024 points/telephone speech, smoothing low	r = **1.0** /0.70
2. 45 Hz — 1024 points/band 0—5 kHz, smoothing low	r = **0.97**/0.41
3. 300 Hz — 1024 points/band telephone	r = **0.96**/0.54
4. 300 Hz — 1024 points/telephone speech	r = **0.96**/0.60
5. 300 Hz — 1024 points/band 0—5 kHz	r = **0.95**/0.45
6. 300 Hz — 1024 points/band telephone	r = **0.94**/0.46

Table III

Phones and words used in study

	300 Hz — 1024 0—5 kHz		**300 Hz — 1024 telephone speech (teleph. band)**		**300 Hz — 1024 simulated telephone band**		**45 Hz — 1024 smoothing none simulated telephone band**		**45 Hz — 1024 smoothing low simulated telephone band**	
phone—word	**same**	**diff**	**same**	**diff**	**same**	**diff**	**same**	**diff**	**same**	**diff**
[a1]—*alla*	1.0	0.65	0.93	0.51	0.96	0.45	0.82	0.35	0.99	0.46
[ll]—*alla*	0.99	0.65	1.0	0.82	0.93	0.83	0.85	0.57	0.99	0.72
[ee]—*teeri*	0.84	0.48	0.91	0.51	0.93	0.57	0.61	0.52	1.0	0.86
[i1]—*pilli*	0.98	0.43	0.98	0.88	1.0	0.69	0.81	0.55	0.95	0.69
[rr]—*parras*	0.97	0.37	0.98	0.59	0.93	0.45	0.63	0.22	0.51	0.28
[mm]—*tamma*	0.88	0.09	0.98	0.23	0.99	0.26	0.68	0.19	0.99	0.27
mean	**0.95**	**0.45**	**0.96**	**0.60**	**0.96**	**0.54**	**0.73**	**0.40**	**0.91**	**0.55**
same/diff ratio (mean values)	**2.11**		**1.60**		**1.78**		**1.83**		**1.65**	

Correlation coefficients (r) of some options are shown in different spectral bands. Column "same" shows the intra-speaker correlations, and column "diff" the inter-speaker correlations. Three snaps were always averaged.

According to the results, the SoundScope FFT option 45 Hz shows the greatest intra-speaker similarity. The first two in the list above consist the telephone band (the first in telephone speech and the second using the telephone band for normally recorded speech). It also seems, that the "smoothing low" option (45 Hz — 1024 points) is far better than the "smoothing none" option. The mean "same" correlation values using the option "smoothing none" were in all cases r = 0.70—0.79.

Within the single phonemes and their correlations (Table III), the simulated telephone band and telephone speech seem to yield a very robust correlation for [a] among vowels (high "same" score and low "diff" score including also the other tables which are not shown in this paper). [mm] shows considerable great differences in all options. The trill [r] seems to be very robust among consonants within both telephone band and 0—5 kHz band. Trills are quite difficult phonemes to measure spectrally because of their open and close phases. It is very important to measure the spectra always at the same location within a period.

5.3. Similarity evaluation of two sisters (Tuija Niemi-Laitinen)

The two female speakers to be compared sound alike (they are in fact sisters; cf. Niemi 1994). Vowels [a, i, o] as well as consonants [s, r, nn] were used in this study (see Table IV below). SoundScope options 300 Hz — 1024 points, 300 Hz — 512 points as well as 45 Hz — 1024 points were used. No smoothing option was used. Correlations between the average spectra were calculated in every spectral band (0—5 kHz, 0—1.5 kHz, telephone band and 1.5—3.4 kHz). Table IV shows an example of these measurements.

Table IV

Phones and words used

300 Hz — 512 points — 3 spectra	**0—5 kHz**		**0—1.5 kHz**		**telephone filter**		**1.5—3.4 kHz**	
phone—word	**same**	**diff**	**same**	**diff**	**same**	**diff**	**same**	**diff**
[a]—*sara*	0.90	0.63	0.96	0.91	0.97	0.62	0.94	0.79
[i]—*siru*	0.90	0.68	1.0	0.99	0.93	0.75	1.0	0.68
[o]—*sokko*	0.73	0.84	0.92	0.83	0.89	0.93	0.73	0.87
[s]—*sara*	0.85	0.94	1.0	0.35	0.92	0.89	0.98	0.94
[r]—*sara*	0.95	0.16	1.0	0.71	0.99	–0.25	0.96	0.84
[nn]—*manna*	0.93	0.85	1.0	0.85	0.98	0.83	0.77	0.70
mean	**0.88**	**0.68**	**0.99**	**0.77**	**0.95**	**0.63**	**0.90**	**0.80**
same/diff ratio (mean values)	**1.29**		**1.29**		**1.51**		**1.13**	

Correlation coefficients (*r*) of one option — 300 Hz, 512 points — in different spectral bands are shown. Column "same" shows the intra-speaker correlations, and column "diff" the inter-speaker correlations. Three snaps were always averaged.

All the mean "same" correlation values (intra-speaker variation) are greater than "diff" correlation values (inter-speaker variation) in all spectral bands. This means that there is more spectral variation between speakers (same utterance) than within one speaker's repetitions of the same utterance. The highest mean "same" values (low intra-speaker variation, mean *r* over 0.90) compared to the mean "diff" values in different spectral bands (including all the SoundScope analysis options used) are as follows:

1. 300 Hz — 512 points/band 0—1.5 kHz	r = **0.99**/0.77
2. 300 Hz — 1024 points/band 0—1.5 kHz	r = **0.96**/0.78
3. 300 Hz —512 points/band telephone	r = **0.95**/0.63
4. 300 Hz — 1024 points/band telephone	r = **0.95**/0.62
5. 300 Hz — 1024 points/band 0—5 kHz	r = **0.92**/0.70
6. 300 Hz — 512 points/band 1.5—3.4 kHz	r = **0.90**/0.80

These results show that the SoundScope FFT option 300 Hz shows the greatest intra-speaker similarity. The first two in the list above compare the lowest channel (0—1.5 kHz), where there is a lot of linguistic similarity (both intra- and inter-speaker similarity). Therefore, telephone band results (300 Hz — 512 and 1024 points, *r* = 0.95) as well as the whole band results (0—5 kHz) show more speaker specific information. There was not much variation in the correlation results between 300 Hz — 1024 points and 300 Hz — 512 points. The correlation values were a little bit smaller with the option 300 Hz — 512 points. Because no smoothing options were used, the correlations using the option 45 Hz — 1024 points were much smaller than the others (both inter- and intra-speaker correlations).

Table IV shows that [a] seems to be the most robust one among vowels in telephone band (*r* same = 0.97 and *r* diff = 0.62). The trill [r] seems to be the most robust among consonants (*r* same = 0.99 and *r* diff = –0.25).

6. Summary

In all three cases, clear spectral differences were observed between two speakers who auditorily sound very similar. The differences in the repetitions of the same linguistic structures produced by the same speaker yielded much lower values.

The following preliminary results can be reported: (1) the effect of the number of the single snaps (3 or 6 snaps within a short period of time) have no significant effect, but the use of one snap only yields better results only in the combination of the smoothing option; (2) the following options yield greater differences between two speakers and a greater similarity in the intra-speaker comparison: a broader filter (300 Hz compared to 45 Hz), greater number of frequency points within the same frequency band, the spectral smoothing; (3) the dissimilarity values get greater, if the band to be compared is larger (e.g. 0—5 kHz and the telephone band); (4) the telephone band and the simulated telephone band yielded very similar results; (5) among different speech sounds (phones), certain sounds or sounds in certain contexts show greater difference between two speakers than the others; (6) the option "filter 45 Hz & smoothing" seems to combine the good effects of a stable spectral form and a relative great independence on the fundamental frequency.

BIBLIOGRAPHY

Altosaar, T., Meister, E. 1995, Speaker Recognition Experiments in Estonian Using Multi-Layer Feed-Forward Neural Nets. — Proceedings of 4th European Conference on Speech Communication and Technology EUROSPEECH'95, vol. 1, Madrid, 333—336.

Harinen, K. 1996, Puhujantunnistus forensisen fonetiikan näkökulmasta — puhelimen vaikutus sekä muut virhelähteet (Master Thesis, Department of Phonetics, University of Helsinki).

Niemi, T. 1994, Puhujantunnistus foneettisena ongelmana (Master Thesis, Department of Phonetics, University of Helsinki).

Nolan, F. 1983, The Phonetic Bases of Speaker Recognition, Cambridge.

Paliwal, K. K. 1983, Effectiveness of Different Vowel Sounds in Automatic Speaker Identification. —Journal of Phonetics, vol.12, 17—21.