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

Tuija Niemi-Laitinen*†, Antti Iivonen*, and Kirsi Harinen*
*Department of Phonetics, University of Helsinki, Finland,
†Crime Laboratory of Finnish National Bureau of Investigation

ABSTRACT

The experiments showed that FFT spectra can be used for the expression of the similarity degree between two speakers. The similarity measure is based on the mean correlation coefficient (Pearson) obtained in a comparison of spectral data representing 12 phonemes. In all comparisons inter-speaker similarity appeared to be smaller than intra-speaker similarity. All technical devices were kept unchanged in the comparison of speaker pairs. Smoothed linear spectra and a 45 Hz broad filter were applied in the FFT analysis. The comparisons made with stressed syllable data showed more inter-speaker similarity than the data in which stress was disregarded. Some phonemes showed more individual character than others. In different contexts the phonemes showed different similarity degrees. Two forensic cases will be discussed.

1. INTRODUCTION

Acoustical argumentation in forensic phonetics can be based on the following properties of speech: voice quality (LTAS, jitter, shimmer, spectral tilt), voice characteristics (including average, range, and standard deviation of fundamental frequency), prosodic features, transitional features and those of speech sounds (phones). Individual differences can be found in all these features. We will concentrate our attention below on the question of to what extent speaker discrimination can be based on spectral snaps from temporal mid-points of speech sounds (phones). We also included plosive bursts in the data. Our point of view is forensic, and this influenced the decisions made in the research procedure. Our main interest is in those speakers who sound very similar.

It can be assumed that a considerable number of the speaker's individual characteristics are included in sound spectra. Because sound spectra are physical correlates to phonological entities (phonemes), it is understandable that spectra of the same phoneme must be to a certain extent similar in two speakers. The additional individual difference is interesting for speaker verification, identification, and discrimination. It would be ideal, if the similarity in repetitions of the same structure were to remain great within the same speaker, but it gets lower, when another speaker produces the same structure. Different types of speech sounds may include more individual variation than the rest. We will discuss this aspect taking into account different Finnish vowel and consonant types (cf. also [1], [2], [3]).

2. RESEARCH HYPOTHESIS

According to our earlier research [2], the short time FFT-spectra of phoneme realizations show greater intra-speaker similarity degrees and greater inter-speaker difference degrees in repetitions of the same linguistic structures in those cases in which the voices of two speakers sound very similar. The spectra were produced on the SoundScope program and processed by our Spectral Comparison program, created for this special purpose. Correlation analysis (Pearson) was applied.

The following hypothesis was established: **Speaker specific features are involved in single speech sound realizations. Under the technical circumstances described below, intra-speaker variation measured by means of spectral comparison is smaller than inter-speaker variation, in spite of the fact that the utterances compared are linguistically identical and the speakers sound very similar.**

Although it can be assumed that in forensic comparisons deviant sound types occur in some speakers, more frequently cases occur in which the individual differences in phones are not clearly observable auditorily. Indiosyncratic differences can appear more clearly in some specific phones, but these kinds of idiosyncrasies can also be speaker dependent.

We have paid attention to the question of how the spectra of phones in Finnish differ in this respect and how stable the phone dependent differences are within the comparisons of speaker pairs. On the basis of preliminary experiments, we concluded that when the average of correlation coefficients of a comparison of 12 phoneme realizations is reached, a saturation level results: the average does not change after adding data.

It is also possible that some parts of the spectrum show more idiosyncrasies than others. Some of the results are reported in [2].

3. ANALYSIS METHODS AND OPTIONS

We used the **SoundScope** speech analysis program (GW Instruments) and our principal option was the short time FFT spectrum (snap). On the basis of our earlier evaluation of the results, we limited the analysis options to the following choices: sampling rate 22050, pre emphasis (6 dB/octave), 1024 points within 11025 Hz, limitation to the telephone band (300-3400 Hz), filter 45 Hz, time window 33 ms (Hamming), medium smoothing (cf. Fig 1), measurement point at temporal mid-point of the speech sound. A combination of a narrow band filter, a relatively long time window and smoothing, yields a stable spectral presentation. In the speech recordings the test arrangements were kept identical in all speech data pairs. 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 imported (copy wave to text) into the **Spectral Comparison** program, based on the FutureBasic II programming language. This program calculates the mean 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 for several single spectra and then to compare two averaged spectra with each other.

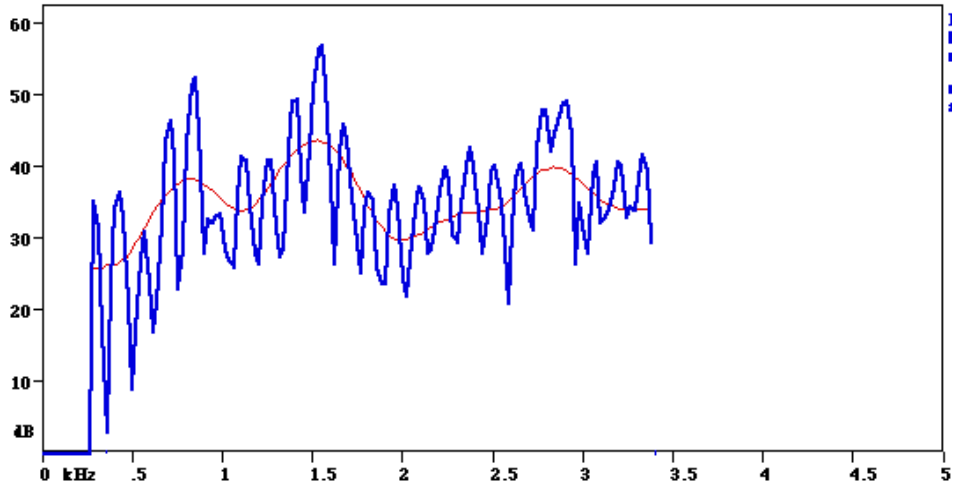

Figure 1. The vowel [a] in the word *anna* produced by a male speaker, and displayed with and without spectral smoothing. The FFT data points were imported from the SoundScope analysis using the following options: sampling rate 22050, pre emphasis (6 dB/octave), 1024 points within 11025 Hz, filter 45 Hz, time window 33 ms (Hamming), medium smoothing, measurement point at the temporal mid-point of the speech sound. The data were further processed by the Spectral Comparison program. This procedure was limited to the telephone band (300-3400 Hz).

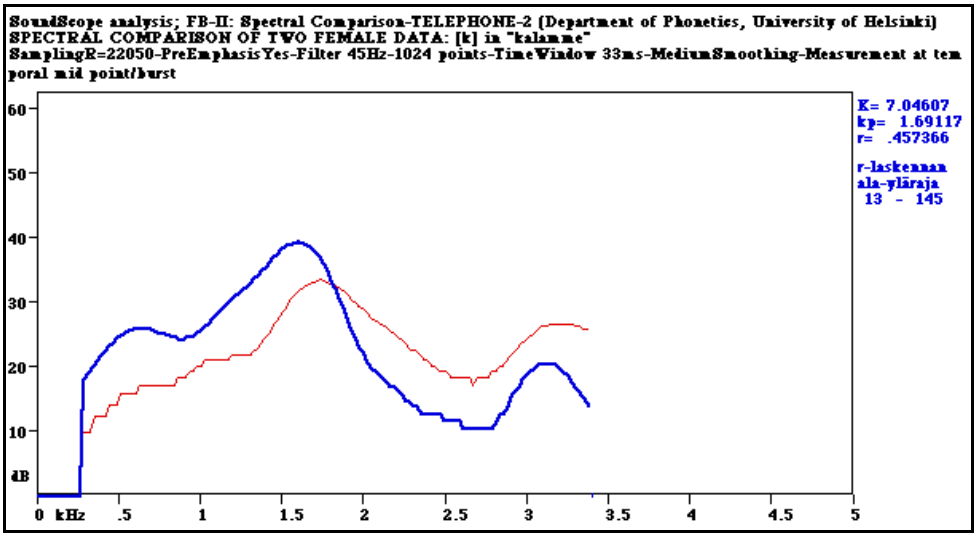

Figure 2. Spectral comparison within the telephone band. The burst phase of the test phone [k] in the word *kalamme* produced by two young female speakers MA (thin curve) and TA (thick). The measurement point statistics on the Hz scale are also indicated on the computer screen: K = mean, kp = mean deviation, r = correlation coefficient (Pearson).

For comparison of two spectra (or two averaged spectra), the program applies the principle of best fit by averaging all the differences at the measurement points on the frequency scale and making one of the two spectra resemble the other by subtracting the average value from all the values of the other spectrum. The two spectra to be compared and the statistical values obtained are plotted on a computer screen.

4. SOURCES OF ERROR

Among other things, the following factors affect the form of the spectrum of a single speech sound: temporal location of the measurement point within the speech sound, surrounding speech sounds (coarticulation), degree of stress, height of fundamental frequency, location within a single period (in resonant sounds), voice quality, emotion, random variation, speech style, type of analysis option, recording circumstances, and recording devices. Several single spectra can be gathered at intervals of 10 ms or at the same pitch synchronous measurement points (single snaps from different periods, but always at the same location within a period) to make from these an averaged spectrum, in order to avoid casual and minute (unimportant) variations [2]. Thus, a more stable spectral form can be obtained. Without this

averaging undesired extra variation will result. A stable picture can also be achieved using a narrow band filter combined with the smoothing option.

5. MATERIAL SELECTIONS

We have used texts, sentences, and isolated words. A part of our material consists of authentic forensic data. We have included phones from stressed and unstressed syllables. In all comparisons the phones to be compared occurred in the same linguistic structures. A part of the material has been recorded through a telephone and another part through a microphone. In all cases, only the telephone band 300-3400 Hz was taken into account in the final spectral analysis.

We took into account the statistical frequency of Finnish phonemes in the sense that the less frequent phonemes were not included in the data. Note that about a half the phonemes occurring in spoken texts in Finnish are vowels and the short and long vowels can occur in all the syllables of a word.

6. COMPARISON OF SPEAKER PAIRS

Our report concerns the following (non-forensic) spoken materials and correlation coefficients. In all comparisons the mean of the correlation coefficients has been calculated from 12 phone pairs.

1) Two male speakers AK and JH (with similar voices and same dialectal background) produced the same text via telephone. The texts were recorded on an answering machine (Panasonic EASA-PHONE).

When the stress position of the phones was disregarded, a mean correlation r (inter) = 0.66 was obtained. The intra-speaker correlation in AK's material was r = 0.82.

2) The same experiment was repeated, but only vowels in stressed syllables and consonants on the border between stressed and a following syllable were regarded. Two sets of 12 phonemes were analysed.

The mean correlations r (inter; AK & JH) = 0.79 and 0.74 were obtained. (Cf. also Tables 2 and 3.)

The mean correlations for the two intra-speaker comparisons were (AK & AK) 0.85 and 0.87.

3) One male speaker MN (41 yrs) produced the same text as at the point 1 text via telephone in December 1995 and repeated it on October 1998. The time span between the recordings was about 3 years. Stress position was disregarded. Mean correlation r (intra) = 0.82.

When the test material was selected as in experiment 2, equal r values were obtained: 0.81 and 0.82 (cf. Tables 2 and 3).

4) Two sisters (ages 31 and 41 yrs) repeated isolated words. Mean correlation coefficients r (intra) = 0.94 and r (inter) = 0.74. The recordings were made using a AKG C567 E microphone and a Marantz cassette recorder.

5) A pair of identical male twins A and E (27 yrs) read the same short sentences. Mean correlation coefficients r (intra) = 0.81 and r (inter) = 0.75. In intra-speaker comparison, the speaker's voice was breathy in the first recording. The sentences were recorded directly on a computer hard disk using a high quality microphone.

6) Two young female speakers MA and TA (19 and 20 yrs) with a similar South Finnish linguistic background produced the same text (a recipe for fish food). The test vowels and consonants were selected from the first (stressed) syllable of the words. The texts were recorded using a AKG-C-451E microphone on a Revox A700 recorder and copied on a TEAC W-440-C cassette recorder. Mean correlation r (inter) = 0.79.

The findings are summarized in Table 1. The table shows that in all comparisons the hypothesis turned out to be correct (in spite of voice difference in one intra-speaker comparison (cf. point 5) and a time span of over two years between the recordings in another comparison (cf. 3). In comparison Nr 6, the inter-speaker correlation was 0.79 indicating a high degree of similarity between the two female speakers.

EXPERIMENT ->	1	2	3	4	5	6
SPEECH MATERIAL	2 males	2 males	2 sisters	identical male twins	a male speaker with 2 yrs time span	2 females
intra-correlation	0.82	-	0.94	0.81	0.82	-
inter-correlation	0.62	0.79 & 0.74	0.74	0.75	-	0.79

Table 1. Intra-speaker and inter-speaker variability expressed as the averages of correlation coefficients of FFT spectral data representing 12 phonemes.

7. FORENSIC CASES

In forensic applications, technical recording and speaker behavior (emotion and style) can differ considerably from ideal test arrangements, and these features can be very different in criminal and suspect data. Our report includes a case in which a female suspect confessed to be the person who made an original false alarm. The technical quality differed in the original false alarm recording and the recording of the same utterance spoken by the suspect under police direction. The auditory impression was that the suspect was the same person as the criminal. Both speakers had very breathy phonation and no voice disguise could be detected. The correlation (mean r = 0.40; 12 phone comparison) remains much lower than in the other intra-speaker comparisons.

In another case (false alarm) the analysis options differed from those reported above: 10 phones were analyzed, the filter breadth was 300 Hz, 3 snaps were averaged for every phone representation. The material produced by a male speaker included so many repeated words that inter-speaker and intra-speaker comparison was possible. In this case, the original spontaneous speech (alarm) and the suspect's read speech were recorded via telephone by means of the same recording device. The correlation coefficients 0.83 (inter) and 0.87 (intra-suspect) were obtained. Because all other speaker characteristics indicated a high degree of similarity between the speakers, the conclusion was that they were very probably the same person.

Note that the Finnish telephone net has been totally digitalized. This technical fact helps reduce spectral variation.

Our conclusion is that the spectral correlation analysis belongs to the appropriate methods contributing to show the degree of similarity between two speakers, if the analysis options are well selected and the recording devices similar.

8. DEGREE OF SIMILARITY OF PHONES IN DIFFERENT CONTEXTS

Tables 2 and 3 show the correlation coefficients in three intra-speaker or inter-speaker comparisons based on a spectral analysis of two different sets of 12 Finnish phones. Only vowels in stressed syllables and consonants on the border between a stressed and a following syllable were included in the comparisons. Otherwise the environment differed. Only phonemes occurring frequently in Finnish are considered.

In the intra-speaker comparisons (AK & AK and MN & MN), the correlation coefficients are mostly very high, but some phones get lower values. These phones (*e*, long *i*, *d*, and *s*) all have high spectral components. The phone *m* is an exception. The mean correlations are equal in both comparisons: r = 0.85 and 0.87 (AK) and 0.81 and 0.82 (MN).

The correlation coefficients in the inter-speaker comparison (AK & JH) were lower for both sets of phones: r = 0.78 and 0.74, but higher than in the experiments in which all stress positions were allowed (cf. Chapter 6, point 1). Many comparisons do show high correlation, but on the other hand, many correlations are much worse than in the intra-speaker comparisons. The latter cases are interesting for forensic applications. This group includes short *a*, long *l*, short *s*, *k* burst.

Coarticulatory effects may explain the fact that some phones can have a high correlation in the first set of phones, but a low correlation in the second (cf. *nn* r = 0.63 in Table 2 but 0.88 in Table 3).

1 phone & word	2 AK&AK	3 MN&MN	4 AK&JH
a1-sadeajan	0.9299	0.7767	0.6907
aa-vaaran	0.8451	0.8748	0.8651
e-keskimäärin	0.6342	0.8669	0.8162
ii-piilossa	0.9631	0.5071	0.9011
o-kovin	0.9302	0.9572	0.9325
u-useimmiten	0.9050	0.8439	0.9349
y -kykenee	0.8591	0.8418	0.7155
ä-tästä	0.8796	0.9783	0.8886
d-sadeajan	0.9275	0.6874	0.9077
ll-sillä	0.9467	0.8262	0.4001
nn-synnyttää	0.9084	0.8227	0.6255
s-keskimäärin	0.5288	0.7777	0.7460
mean	**0.8546**	**0.8133**	**0.7853**

Table 2. Correlation coefficients in an intra-speaker and inter-speaker comparison based on a spectral analysis of 12 Finnish phones. (1) test phone and word, (2) intra-speaker comparison (speaker AK), (3) intra-speaker comparison (speaker MN with about 3 years time span), (4) inter-speaker comparison (speakers AK and JH).

1 phone & word	2 AK&AK	3 MN&MN	4 AK&JH
a-sarvikuonon	0.9280	0.7687	0.6640
e-selittyy	0.9542	0.9273	0.7293
ii-liikkumaan	0.8760	0.6657	0.8250
o-prosenttia	0.9426	0.9456	0.8917
uu-suunnasta	0.9116	0.9384	0.8788
y-pysyttelee	0.9288	0.8733	0.7189
ä1-säpsähtää	0.9300	0.9378	0.8705
nn-suunnasta	0.7330	0.9177	0.8782
s-pysyttelee	0.6040	0.2950	0.4526
r-sarvettoman	0.8773	0.9267	0.8695
k2-kykenee	0.9210	0.9004	0.2600
m-samantapaista	0.9377	0.6849	0.7927
mean	**0.8787**	**0.8151**	**0.7359**

Table 3. Correlation coefficients in an intra-speaker and inter-speaker comparison based on another set of 12 Finnish phones. (1) test phone and word, (2) intra-speaker comparison (speaker AK), (3) intra-speaker comparison (speaker MN with about 3 years time span), (4) inter-speaker comparison (speakers AK and JH).

9. CONCLUSIONS

Our results show that a correlation analysis of FFT spectra can be used as a measure of the similarity degree between speakers who sound very similar. The speech sounds include individual properties which can be revealed using FFT analysis. The results were achieved by limiting the analysis band to the telephone band 300-3400 Hz. The forensic applications appear promising.

If speaking style, speaker's emotional state and recording devices change in the criminal and suspect data, the correlation coefficients worsen.

Some speech sound types show more individual differences than others and are therefore better candidates for forensic comparison. The phonemes behave in different ways in different contexts and hence the forensic value of the contexts varies. A larger investigation of these aspects will be needed.

ACKNOWLEDGMENTS

We would like to thank Leena Keinaenen for the twin material and the Crime Laboratory of the Finnish National Bureau of Investigation for the forensic data.

REFERENCES

[1] Nolan, F. 1983. *The Phonetic Bases of Speaker Recognition.* – Cambridge University Press.

[2] Iivonen, A., Niemi-Laitinen, T. & Harinen, K. 1998. Evaluation of similarity degree between speakers on the basis of short time FFT spectra. Proceedings of the Finnic Phonetics Symposium August 11-14, 1998, Paernu, Estonia. *Linguistica Uralica*, XXXIV/3, 192 –198.

[3] Paliwal, K.K. 1983, Effectiveness of different vowel sounds in automatic speaker identification. *Journal of Phonetics*, 12, 17–21.