https://doi.org/10.3176/lu.1998.3.05

TUIJA NIEMI-LAITINEN, PÄIVIKKI ESKELINEN-RÖNKÄ (Helsinki),
RISTO MONTO (Vantaa)

SPEAKER DATABASE TEST AND FUNDAMENTAL FREQUENCY IN SPEECH

This is a study of a speaker database program and fundamental frequency in speech in Finnish. To determine the capability of the database program used, a set of test runs was initiated as a co-operation project of the Department of Phonetics (University of Helsinki) and the National Bureau of Investigation (NBI). When collecting the voice samples of over one hundred speakers, an attempt was made to simulate an authentic recording session as accurately as possible, both from a technical standpoint and in terms of sample taking. NBI standards were applied to the whole testing procedure.

The results of a closed test set indicate the adequacy of the program when original and reference samples came from the same channel. In a test with no search restriction criteria in force, the program managed to identify a same-channel speaker in 90% of the cases. When the search was restricted on the basis of gender the rate of identification was 92% for male speakers and 94% for female speakers. When the channel of reference samples were changed into a simulated telephone channel a dramatic drop in the identification rate was observed: In a test with no search restriction in force the program was successful only in 32% of the cases and with restricted gender-based search the identification rate for male speakers was 38% and for female speakers 36%.

1. Fundamental frequency in speech

1.1. F0 analysis

The fundamental frequency analysis has been quite successfully used in forensic phonetics (see for exampie LaRiviere 1975; Braun 1992; 1995; Gfroerer, Wagner 1995; Ladd, Terken 1995; Niemi 1995; Jessen 1997). The fundamental frequency in speech is dependent on the anatomy of the speaker, but it is quite easy to disguise, too. Voiced sounds and sections of speech are needed for F0 analysis. The Finnish language has many vowels and diphthongs, which form the bases of this analysis. Fundamental frequency in speech is resistant to channel distortions and it is measurable even from telephone transmitted speech. The fundamental frequency in speech is sensitive to the following distortions: the age of the speaker, alcohol, drugs, the emotional state of the speaker, speaking style and situation, duration of the speech sample, tape speed as well as noise in the area of F0.

Some problems may arise during the F0 analysis. First, the analysis range needs

to be set. This means that the lowest and highest frequency of a certain speaker have to be anticipated. Something very typical for the speaker may be missed if too narrow a range is used. The other problem arises with different algorithms. For example, Medav Spektro 3000 has two different algorithms for F0 analysis: cepstrum and SIFT (Simplified Inverse Filter Tracking). From these two, cepstrum uses spectrogram and fundamental frequency harmonics to calculate the F0. It is then useful when telephone speech needs to be analyzed. SIFT uses inverse filtering to measure the F0. These two methods are explained below.

1.2. Cepstrum

The spectra of the color spectrograms are the starting point for this algorithm. The dynamic limit needs to be carefully set for successful analysis. The individual spectra are limited to 2 kHz and subjected to an FFT again yielding the cepstrum showing the harmonic components of the spectrum i.e. the fundamental voice frequency (pitch) with its harmonics. Even when the fundamental is not alone in the signal (e.g. telephone transmission) it can be seen by the harmonic spacing. The cepstrum's maximum is finally converted to the pitch value. For small signal sections, (color spectrum overlapping > 50%) a pitch value is calculated for each individual spectrum of the color spectrogram and displayed in a polyline diagram. Outliers are then removed by third-order median filtering (each value has its predecessor and follower added to itself, the largest and smallest values are removed from the three and the original value is replaced by the remaining one). The curve is finally smoothed out by replacing each value by the medium value of its adjacent values (n corresponds to 40 msec). The polyline consists of pitch minimums and maximums of 50% overlapping time sections for large sections.

1.3. SIFT algorithm

The SIFT (Simplified Inverse Filter Tracking) algorithm in Medav Spektro determines the fundamental voice frequence via a maximum in the ACF of the LPC residual signal. The speech signal is split up into segments of 16—100 milliseconds in length, and each one has its pitch calculated. The signal bandwidth is first limited to 1 kHz, using low pass filtering. This helps to differentiate between the presence or absence of a voice signal by suppressing high frequency components such as hiss and explosive sounds and thereby removing energy from these sections. Most of the energy of a voice signal is below 1 kHz. A Durbin Levinson recursion uses the first five ACF values of each segment to calculate the four coefficients of an inverse filter used for reversing the speaker's speech tract modulation. Ideally, inverse filtering the segment results in a residual signal of white noise or a pulse train. This can be interpreted as the speech tract's excitation signal.

The pitch results of the individual segments are displayed on a polyline diagram. Voiceless sections cause the line to be positioned on the lower edge of the diagram. Outliers in the voice sections are then removed by 3rd-order median filtering.

2. Speakers and speech material

The fundamental frequency of 111 Finnish speakers (53 women, 58 men) was analyzed. Female speakers were 21—50 years old and male speakers 20—40 years old (see Table 1 below). The same speech material was also used in testing a speaker database program explained in section 5. For more details in recordings and speakers, see

also Eskelinen-Rönkä 1997 : 48—57. The speaking material consists of a reading passage of about 20—40 seconds, depending on the speaking rate. The passage is as follows:

Lähes kaikki, mitä nykyihmisellä on, on rahalla ostettavissa. Hänen yksityisetkin asiansa ovat useimmiten samanlaisia kuin monilla muilla. Niihin ei kätkeydy persoonallista salaisuutta, sillä ne on hankittu kauppakeskuksista, tavarataloista, supermarketeista, kirpputoreilta tai kioskeista. Ostoksilla käyminen on hänelle yksinkertaisesti jo elämän ylläpidon kannalta välttämätöntä, sillä vaihdon yhteiskunnassa vain äärimmäisen harva on omavarainen.

Table 1

Data on speakers and speech material used in F0 analysis

	MALE	FEMALE
Number of speakers	58	53
DIALECTAL VARIATION		
Standard Finnish	23	24
Finnish + Swedish	5	6
Dialect speakers	30	23
Smokers/non-smokers	14/44	13/40
AGE:	years	years
Mean	24.3	30.6
Min	20	21
Max	40	50
SPEECH DURATION	seconds	seconds
Mean	29.2	30.0
Min	23.3	23.7
Max	36.8	44.8

3. Analyses

Many parameters may be calculated while analyzing the fundamental frequency in speech, for example the mean or the average, the standard deviation and the range. At least 20—30 seconds of speech is needed to calculate the average fundamental frequency in speech (Nolan 1983 : 121—130, and own research). Very little variation is found after this point. The standard deviation of fundamental frequency tells how much variation there is from the average value (+–). The range value can be expressed either in hertz (Hz) or semitones (ST). This range expresses the lowest and the highest F0 value that a certain speaker uses while speaking. If a speaker is said to have very lively voice, it means that this speaker has a wide range of F0. Monotonous speakers, on the other hand, have a very narrow range of F0.

In this study we analyzed mean fundamental frequency (F0) and its standard deviation (STDEV). Both cepstrum and SIFT algorithms were used. With SIFT we used three different options: 1) a segment length of either 16 milliseconds (with 1% threshold), 2) 30 milliseconds (with 10% threshold) or 3) 50 milliseconds (with 20% threshold).

4. Results

The results of F0 analyses are shown below in Table 2 (female speakers) and Table 3 (male speakers). In both tables the first column is for the F0 average, the second is for the F0 standard deviation. These two columns are first shown for cepstrum results and then for SIFT results.

Table 2

Mean values of F0 and its standard deviation for female speakers (N = 53)

cepstrum mean (Hz)	stdev (Hz)	SIFT 16 mean (Hz)	stdev (Hz)	SIFT 30 mean (Hz)	stdev (Hz)	SIFT 50 mean (Hz)	stdev (Hz)
182.3	20.1	183.2	22.5	186.2	23.3	187.6	24.1

As can be seen, cepstrum value is the minimum and SIFT 50 value the maximum (see text for explanations).

Table 3

Mean values of F0 and its standard deviation for male speakers (N = 58)

cepstrum mean (Hz)	stdev (Hz)	SIFT 16 mean (Hz)	stdev (Hz)	SIFT 30 mean (Hz)	stdev (Hz)	SIFT 50 mean (Hz)	stdev (Hz)
105.9	10.1	102.7	11.2	101.7	11.6	102.6	11.4

As can be seen, cepstrum value is the minimum and SIFT 30 value the maximum (see text for explanations).

The results show that with male speakers, the cepstrum values of mean F0 are the greatest of all and SIFT 30 values the smallest (105.9 vs. 101.7 Hz). Results of female speakers, on the other hand, show that SIFT 50 are the greatest and cepstrum values are the smallest (187.6 vs 182.3 Hz). For male speakers the 30 ms segment length with SIFT algorithm should be the most reliable one (inverse filtering needs to be set so that frame length is at least twice the speakers F0: if F0 is about 100 Hz, then the frame length should be at least 20 ms). With female speakers the frame length should be shorter than with male speakers, so it is obvious that SIFT 16 (16 ms frame length) is the most optimal one for female speakers. This SIFT value is indeed the nearest one to cepstrum value (183.2 vs 182.3 Hz).

There are, however great differences between the mean values of these algorithms. When, for example, SIFT is not set for the optimal analysis options, it does not measure all the voiced sections of speech. In forensic cases, it is important to guarantee that telephone transmitted speech and its F0 can both be analysed. Thus, we recommend the use of the cepstrum algorithm, on the understanding that the signal is amplified so that the program finds every voiced part of the speech.

5. Speaker database research

In 1989, Finland legalized the recording and registering of sound samples of suspects so that they could be used as distinguishing markers. After this new law went into effect, NBI began developing a database program for voice registers. Here the objective was to acquire a database application that was conducted reliably and efficiently and that would allow the registration of speech samples, their comparison, as well as the comparison of speech samples in and outside the register. The Crime Laboratory of NBI received the first prototype version of an automatic database program in 1995.

The database program that is currently in use is based on the metrification and vector quantification of the parameter values. The parameters computed from the speaker's speech sample to be registered are vector quantified into code books about the speaker. These code books are recorded in the database with other speaker-specific data. During comparisons or searches the speech sample of an unknown speaker is parametrified and quantified without the teaching of code books, which is computationally a complicated operation. The code vectors which are established are compared with the code vectors of the code books included in the database register. The comparison of code vectors can either be extended to cover the whole data-

base, it can be restricted to a particular crime type, or it can cover only one gender, so that the search can be limited to a particular speaker group. As a result of the comparisons, the program provides, according to the restrictive criteria applied, the personal data of the speakers in the register that best match the speech sample of the speaker being compared. The rate of match between the speech samples and the sample of the speaker being compared is expressed as a distance ratio.

The program has been pre-tested by the manufacturer and the buyer. Both tests were conducted with identical parameter settings using an equal size, 20-speaker data base. Test results gave a very optimistic picture of the program's capacity and capabilities. However, when we take into account the planned purpose and environment of the program, small test runs conducted with restricted and clearly distinct speaker data do not give a full picture of the capability of the system to identify speakers. When the system is used by the police as a voice register application, it has to provide reliable results based on speech data from hundreds, sometimes thousands of speakers.

To determine the real capability of the database program, a larger set of test runs was initiated in a joint Phonetics Department and NBI project. To maintain the comparability of all test results referring to the same system, the optimized parameter settings already programmed in were not changed. When collecting the voice samples of over one hundred speakers, an attempt was made to simulate an authentic recording session as accurately as possible, both from a technical standpoint and in terms of sample taking. NBI standards were applied to the whole testing procedure in order to identify and eliminate the error sources that affect methodology.

The database program was taught with one hundred (C-cassette linear recording) speech samples (50 male, 50 female). The length of each sample was five seconds and it consisted of personal information on each speaker, such as name and address. All of the samples were edited in order to remove any additional sounds or silences that might have a negative affect on the program, e.g. pauses while speaking and breathing sounds. As a reference sample a five-second speech sample was taken from a reading passage of each speaker (see section 2) and was edited according to the same principles as mentioned above. These edited reference samples were also filtered with simulated telephone channels. The two sample groups used as references in the test run consisted of one hundred same channel (C-cassette linear recording) and one hundred different channel (simulated telephone filtering) speech samples.

The efficiency of the data base program was investigated in a closed test with the two previously-mentioned groups of references. The efficiency of restriction criteria was also tested. The restriction criteria used was based on the gender of the speaker. The results of the basic test run are shown in table 4 (below).

Table 4

Results of the basic test run (closed test)

SEARCH RESTRICTIONS	Total N of samples	Success %
1) NO RESTRICTIONS		
a) same channel references	100	90
b) different channel references	100	32
2) WITH SEARCH RESTRICTION		
a) same channel references		
male	50	92
female	50	94
b) different channel references		
male	50	38
female	50	36

6. Conclusion

It is naturally too early to assess the final potential of the program on the basis of these limited test results. The program will be widely tested both by means of the aforementioned methodology and by other testing methods, and the influence of error sources will also be closely examined. It should also be useful to test how the results of the database tests are affected by the average fundamental frequency in speech.

REFERENCES

B r a u n, A. 1992, Zur Bedeutung des Merkmals "mittlere Sprechstimmlage" in der Forensischen Phonetik in Phonetik und Dialektologie. — Joachim Göschel zum 60. Geburtstag, Marburg (Schriftenreihe der Universitätsbibliothek), 1—26.

— —1995, Fundamental Frequency — How Speaker-Specific Is It? — Studies in Forensic Phonetics, Trier, 9—23.

E s k e l i n e n-R ö n k ä, P. 1997, Raportti automaattisen Puhujan Tunnistaja-tietokantaohjelman testauksesta. Master Thesis, Helsinki.

G f r o e r e r, S., W a g n e r, I. 1995, Fundamental Frequency in Forensic Speech Samples. — Studies in Forensic Phonetics, Trier, 41—48.

L a d d, D. R., T e r k e n, J. 1995, Modelling Intra- and Inter-Speaker Pitch Range Variation. — Proceedings of the XIIIth International Congress of Phonetic Sciences, vol. 2, Stockholm, 386—389.

L a R i v i e r e C. 1975, Contributions of Fundamental Frequency and Formant Frequencies to Speaker Identification. — Phonetica 31, 185—197.

J e s s e n, M. 1997, Speaker-Specific Information in Voice-Quality Parameters. — Forensic Linguistics. Speech, Language and the Law, vol. 4, Birmingham, 84—103.

N i e m i, T. 1995, Puhujantunnistus foneettisena ongelmana — Papers from the 19th Meeting of Finnish Phoneticians, Tampere (Department of Finnish Language and General Linguistics, University of Tampere. Folia Fennistica & Linguistica 18), 101—116.

N o l a n, F. 1983, The Phonetic Bases of Speaker Recognition, Cambridge.