Development of a Multiparametric Speaker Profile for Speaker Recognition

Antti Iivonen, Kirsi Harinen, Leena Keinänen, Jussi Kirjavainen, Einar Meister† and Launo Tuuri
Department of Phonetics, University of Helsinki
† Technical University, Institute of Cybernetics, Tallinn
E-mail: antti.iivonen@helsinki.fi, kirsi.harinen@helsinki.fi, leena.keinanen@helsinki.fi,
jussi.kirjavainen@helsinki.fi, einar@ioc.ee, launo.tuuri@helsinki.fi

ABSTRACT

This report includes (1) recording and storing of different Finnish speech materials including a mobile telephone (GSM) speech data base (about 240 speakers), (2) critical composition of an acoustical multiparametric speaker profile, (3) development of a speaker identification program PROFMATCH based on the use of individual profiles, and (4) clustering and identification tests with hi-fi, medium and GSM quality speech. An automatic multiparametric Praat script **SpeakerProfiler** has been created. An identification score 100% was achieved with 110 speakers in the tests with medium technical speech quality utilising Gaussian Mixture Model and only few basic parameters. Identification of GSM quality was successful, if the recording channel was unchanged. More tests are needed for an improvement of the recognition score, if variable recording channels are applied. Attempts with an automatic segmentation program (SEGMENTER) yielding an access to individual sound segments have been promising.

1. INTRODUCTION

Already in 1963 Garvin and Ladefoged [3] made a differentiation of the two basic types of information which are contained in the voice signal: "message identification" and "speaker identification". In the present study, the basic target was to create a multiparametric speaker profile which utilizes such acoustical features which contribute to an expression of individual speaker properties and which can be used in practical applications. "Practical" means that the identification should be possible also in everyday technical circumstances, e.g. when the transmission takes place via mobile telephone. The applications include speaker discrimination and verification as well as forensic speaker identification based on a speaker profile database.

Critical research was concentrated on the following aspects: technical recording and transmission of the speech material, robustness and discrimination power of the acoustical parameters, minimal time of the speech sample needed (temporal saturation), effect of speech style (read and spontaneous), analysis options and reliability of measurements. Discrimination of identical twin speakers was taken into account.

Mainly the Praat program (developed by P. Boersma and D. Weenink) was used, but some basic work was made by means of SoundScope and Kay CSL programs. A fully automatic Praat script was developed for the creation of the speaker profiles, but attention was also paid to further parameters which must be partly processed manually at the moment.

The statistical analysis package SPSS was used in hierarchical cluster analysis of speaker profiles. The programs DISCRTEST and PROFMATCH were used for parameter evaluation and speaker idenfication (both developed by our partner group at the University of Joensuu; cf. below).

2. SPEECH MATERIAL

The speech samples include a text consisting of 26 sentences as well as spontaneous answers to 12 questions concerning a holiday trip, a spontaneous description of a picture, reading of the Finnish alphabet, a set of numbers and 18 phonetically rich sentences including loan phonemes. A short utterance "Täällä Ninni on purrut hammasta" [literally: "Here Ninni has bitten her teeth."] was created for experiments with speaker verfication and it was read twice. The test utterance includes (phonologically long) [æ:], nasals and liquids which are shown to be interindividually variable [2, 3]. The read text also included a passage which was read twice. A specific material was planned for comparisons of identical twins.

ISBN 1-876346-48-5

3. SPEAKERS AND RECORDING

Three different recording and transmission channels were applied:

1. HI-FI level: high quality equipment was used for recording of 6 males and 6 females. All informants recorded the material described in chapter 2.

2. MEDIUM QUALITY: A microphone and C-cassette recordings were applied for 55 males and 55 females who read a text. An older database with 55 males and 55 females was available.

3. GSM speech: The sending and the receiving GSM phone were Nokia 3330. The GSM speech samples were recorded either a) into a portable computer (Hewlett Packard XE3) attached to the receiving mobile phone (sampling rate 44,1 kHz, 16 bit, linear WAV) or b) into a server computer (sampling rate 8 kHz, 8 bit a-law , transfer via FTP).

The 218 speakers (112 males and 106 females) with the age range 16–65 years represent the main dialectal areas in Finland. In addition, a group of Russian and Estonian subjects speaking Finnish were recorded.

Two pairs of identical male twins were recorded, too, in order to compare their differentiation from each other and from the speakers of a control group.

4. CREATION OF A SPEAKER PROFILE

The structure of the speaker profile development is illustrated in Fig. 1. The programs Praat, SoundScope and Kay CSL were utilized for the acoustical basic work. For Praat, scripts were written for automatic limiting of speaker's F0 range, for automatic capturing of FFT spectra from selected individual sound segments, and for measurement of more usual parameters available in Praat. All sub-scripts were combined to a single **SpeakerProfiler** script.

The following critical aspects were studied: text and style dependence, robustness, time saturation of the parameters, analysis program, analysis options, the influence of the technical sources of error.

The discrimination force of the single parameters was studied using a special program (DISCRTEST) and applying the clustering option by the statistical SPSS program.

By means of a semiautomatic program the speech samples were forced to segmentation according to the previously known text by means of a program called SEGMENTER. It yields a Praat text-grid with the following three tiers: diphones, segments and words. Tests with GSM speech showed great accuracy in segmentation, if it was based on a correct text (Finnish orthographic transcription). Yet, deviations of 10–15 ms from the optimal segment boundaries are possible and some mistakes with longer deviation time might occur, too. Selected phoneme types can be chosen for a detailed analysis, e.g. for measurement of FFT spectra. In addition, temporal prosodical patterns can be included in the recognition process. The attempts to use the automatic segmentation seem to be promising. According to [2], adding phoneme based information improves the recognition score.

5. EFFECTIVE PARAMETERS

The features of segments, prosody and voice quality contribute to the individual speech characteristics (Fig. 1). Two derivatives of other parameters, i.e. LTAS and cepstrum have been proved to be effective and robust in speaker recognition [1]. In practical applications, including recognition of GSM speech, many important parameters suffer from serious drawbacks: their automatic measurement is unreliable or impossible and some of them (parameters of voice quality) are cut away outside the 300–3400 Hz band. However, voice quality leaves some traces in spectral properties of the telephone band and they can be processed in speaker recognition.

The effectiveness of the parameters was tested by the SPSS program and DISCRTEST program. In our experiments the combination of mel-filtered cepstrum coefficients (MFCC), linear frequency cepstrum coefficients (LFCC), long time average spectra (LTAS) and F0 statistics has proven to be effective. The test indicated that only a set of 20 averaged MFCCs are sufficient to a succesful recognition. A fully reliable measurement of vowel formants seemed to be difficult and different measurement options yielded variable formant values. Averaged short time FFT spectra from selected frequently occurring and most interindividually variable sound segments, e.g. [a] and [r], can be measured, but the time window causes a problem: a short time window yields random variation; a long window includes harmonics in vocalic sounds. A 30 ms long window combined with a smoothing option has been tested.

Only few seconds sample is needed for the time saturation of cepstrum and LTAS whereas saturation of F0 average needs a much longer time and the average F0 fluctuates according to utterance types and utterance length. The statistical F0 values median and mode seem to be more robust.

6. IDENTIFICATION PROGRAM

All single speaker profiles are calculated and stored in the profile data base. A new profile is calculated for an

ISBN 1-876346-48-5

unknown speaker and the recognition program PROFMATCH (made using the ANSI-C-language and Gaussian Mixture Model) compares it with the stored profiles, picks up the most similar profiles (their number can be selected) from the profile database and calculates a probability coefficient indicating the score of possible identity with the unknown speaker. Each parameter of parameter set can be given its weight according to its discrimination power. The confidence value of the recognition decision can be indicated. [A separate publication will describe more details of the program.]

7. EFFECTIVENESS OF RECOGNITION

Datafusion (multiparametric approach) turned out to be successful. It yielded a better confidence than the single MFCC or LFCC vectors alone. LTAS turned to have the best confidence, because it matched with the correct speaker with the greatest distance compared to the second best match.

The recognition scores have been presented in Table 1. The recognition of the HI-FI and medium recording level MFCC was successful alone. F0 statistics alone were not very effective.

The clustering with the SPSS program yielded 100% recognition score with the 12 speakers (under HI-FI recording quality).

The identification with PROFMATCH program was perfect (100% correct) with 110 speakers (under medium recording quality).

The tests with the GSM quality showed poorer quality in the identification of 156 speakers (62.8% correct), if the two GSM recording conditions (cf. Chapter 3; 3) were mixed. If only the first condition (recording with the portable HP XE3) with 99 speakers was applied, a recognition score 98.9% was achieved.

Separate tests with the two identical male twin pairs showed that both produced interindividual differences which override the intraindividual variation. But in the both cases the differences compared to the control group were greater than those to the twin brother.

8. CONCLUSIONS

Three technical qualities were used for three speech databases: HI-FI (12 speakers), medium (110) and GSM (218) level. A critically evaluated multiparametric speaker profile was created which can automatically be generated by a Praat script (SpeakerProfiler). A speaker idenfication program PROFMATCH was created which yields the best match for an unknown speaker and a confidence value of the recognition decision. Idenfication tests showed that a perfect identification (100%) can be achieved in few seconds (depending on computer quality), if HI-FI or medium quality recording technique is utilized. So far, GSM quality is more problematic. If the recording channel was unchanged, however, almost 100% score was achieved also with the GSM quality. The speaker profile can further be developed adding the segment level information.

The results predict that speaker verfication tasks will succeed perfectly in circumstances in which microphones without telephone connection can be applied, because good technical quality and cooperative speaker can be combined. Speaker identification tasks via GSM speech need more research.

ACKNOWLEGMENTS

Thanks to the other members of our team (Department of Phonetics, University of Helsinki): Päivikki Eskelinen-Rönkä, Mari Horppila, Hanna Liisanantti, Tuija Niemi-Laitinen, Leena Perälä, Liisa Vilhunen, Olli Rissanen; as well as to the experts of program development: Pasi Fränti, Tomi Kinnunen, Ville Hautamäki and Teemu Kilpeläinen from the Department of Computer Science (University of Joensuu) and Antti Suni (Department of General Linguistics, University of Helsinki).
The work was carried out under the project "The Joint Project Finnish Speech Technology" supported by the National Technology Agency (agreements 40285/00, 40406/01, 40238/02) and titled "Speaker Recognition" (University of Helsinki ProjNr 460325). The project has also been supported by Civil Aviation Administration in Finland, The Finnish Defence Forces/Viestikoelaitos, Finnish National Bureau of Investigation, Accident Investigation Board Finland and Scando OY.

REFERENCES

[1] Rose, P. (2002) *Forensic Speaker Identification*. Forensic Science Series. London/New York: Taylor & Francis

[2] Faltlhauser, R. and Ruske, G. (2001) "Improving speaker recognition performance using phonetically structured Gaussian Mixture Models". *Eurospeech 2001 – Scandinavia*, B26, 751.

[3] Garvin, P.L. and Ladefoged, P. (1963) Speaker idenfication and message identification in speech recognition. *Phonetica* **9**, pp. 193–199.

[4] Nolan, F. (1983) *The Phonetic Bases of Speaker Recognition.* Cambridge/London/etc.: Cambridge University Press.

 ISBN 1-876346-48-5

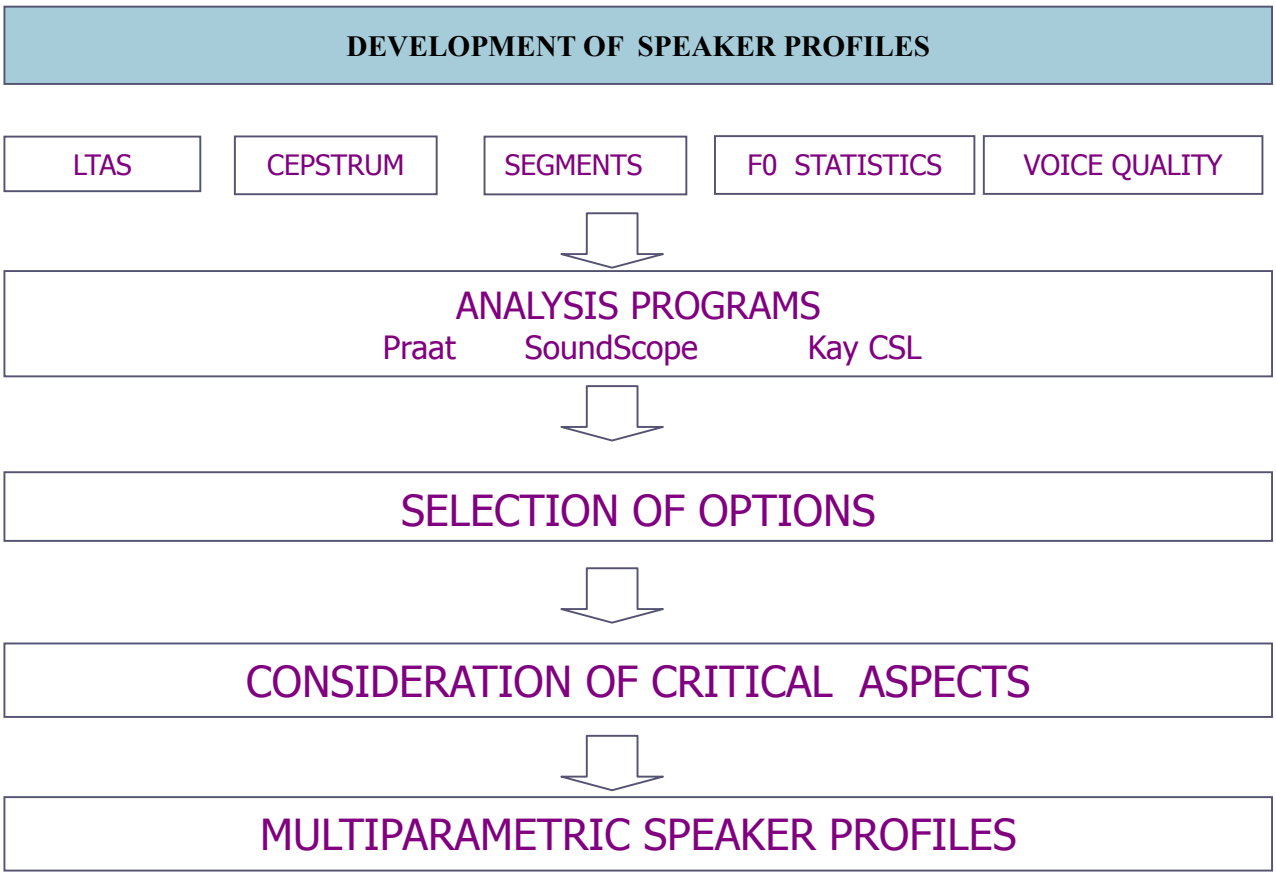

Fig. 1. A scheme for development of speaker profiles. The discrimination power of different acoustic parameters are tested utilizing SoundScope, Kay CSL and (mainly) Praat. The effect of analysis options is tested. Several critical aspects are tested. A fully automatic Praat script has been developed for creation of multiparametric profiles.

Table 1. Effectiveness of speaker identification under some conditions. M=male, F=female. Channel, cf. Chapter 3, recording conditions.

Speech material	**Speakers**	**Duration of speech test sample**	**Recording technique**	**Recognition score**
Short verification utterance (cf, Ch. 2)	6 M 6 F	ca. 2 s	HI-FI	100 %
Read text	6 M 6 F	10 s	HI-FI	100 %
Spontaneous speech	6 M 6 F	10 s	HI-FI	100%
Read text	55 M 55 F	10 s	mikrophone/C-cassette	100 %
Read text against spontaneous speech	157	10 s	GSM Nokia 3330 mixed channel condition	62.8%
Two samples from read speech	99	10 s	GSM only one channel	98.9%

ISBN 1-876346-48-5