APPLYING MFCC-BASED AUTOMATIC SPEAKER RECOGNITION TO GSM AND FORENSIC DATA

Tuija Niemi-Laitinen*, Juhani Saastamoinen, Tomi Kinnunen**, Pasi Fränti****
*Crime Laboratory, NBI, Finland
**Dept. of Computer Science, Univ. of Joensuu, Finland

Abstract

Speaker Profiler computer program for automatic speaker recognition has been developed in a research project funded by the Finnish Technology Agency. A vector quantization (VQ) matching approach is used, where dissimilarity of an unknown speech sample is computed for codebooks created using the K-means algorithm. This study tests the recognition reliability with two databases constructed from Finnish band-limited GSM speech and authentic crime case speech.

Material for the first test is recorded with a GSM phone and a laptop computer. Spontaneous speech vs. reading was tested. The program should pick the right person from the database based on independent non-verbatim speech samples. There were 47.5 % out of 107 samples ranked first correctly. Some very poor quality speech files were used in training and the mother tongue for some speakers was not Finnish. If these samples were not considered, the result was better.

The second part of this study consists of real crime investigation cases. The speaker database was constructed from known speech samples (suspect). Unknown sample(s) recorded at the crime scene were matched against the database. From the matched 61 samples, 68.9 % were ranked first correctly. Accuracy is sufficient for creating shortlists in forensics.

Keywords: speaker recognition, forensics, automatic recognition, real-time recognition, MFCC

1. The aim of the study

The aim of the study is to test how reliably an automatic speaker recognition program performs when the input speech is band-limited (GSM speech) and authentic (real crime case material). Test situation where speech files have been recorded in good laboratory conditions with high quality microphones is far from the reality in forensics. This is why GSM phone recordings were used for this study. Annually most speaker identification cases at the Crime Laboratory consist of GSM speech material only. The tests simulate a typical speaker recognition case. Usually the speech of a criminal (the unknown sample) is spontaneous. The speech of the suspect, on the other hand, usually consists of both reading and semi-spontaneous speech.

2. Speech material and recordings

Table 1 lists some statistics describing the overall structure of the data sets used in this study. These figures are explained in detail in sections 2.1 and 2.2.

Table 1. The number of samples and their duration in GSM and Forensic data

Data		Number of speech samples			Sample duration (sec.)		
Test set	*Subset*	Male	Female	Total	Min.	Max.	Avg.
GSM	Read	47	60	107	140	252	183
GSM	Spont.	47	60	107	48	300	153
GSM	Both together	94	120	214	48	300	168
Forensic	Suspect	27	1	28	6	189	73
Forensic	Crime scene	59	2	61	2.4	379	50
Forensic	Both together	86	3	89	2.4	379	57

2.1. GSM data

Recordings for the first part of the tests were collected during 2001 under the project "The Joint Project Finnish Speech Technology" supported by the National Technology Agency (TEKES agreements 40285/00, 40406/01, 40238/02) and titled "Speaker Recognition" (University of Helsinki Project No. 460325).

For the present study, 107 speakers (60 female and 47 male) were recorded using a GSM phone and portable computer were selected. The samples included 16 noisy recordings as well as 12 samples spoken by persons whose mother tongue was Estonian, Russian or Swedish. The duration of the samples in spontaneous speech varied from 48 to 300 seconds, and in text reading task from 140 to 252 seconds.

2.2. Authentic crime case data

Real crime investigation speech data is used in the second part of the study. The known and unknown speech samples in several crime investigation cases were recorded during the years 2000-2004 either via GSM or land-line phones. The phone type could not be limited to one, because the criminals use the phones they have. One speaker (3 samples) is a female, the others are male. The female has very low-pitched voice, and she was included for testing reasons. Only speech files not containing extra background noise, were considered in the test. Some of the files did still contain some noise, e.g. computer hum and traffic sounds. Total of 61 unknown and 28 known samples were used. The durations of the samples varied from few seconds to 379 seconds.

3. Winsprofiler speaker recognition software

The *Speaker Profiler* (*sprofiler*) is a portable software engine for creating, managing, and recognising voice profiles based on speech samples. The *Winsprofiler* program used in the tests of this study, is a realization of sprofiler that is augmented by a Windows graphical user interface. It supports both training and recognition from sound files or directly from PC-microphone, as well as drag-and-drop input of files. Recognition is fully automated, and the software supports both offline matching from previously recorded samples, as well as real-time matching with speech input stream from microphone of PC-soundcard. The software was developed in a research project funded by the Finnish Technology Agency (TEKES agreements 40437/03 and 40398/04).

Speaker Profiler uses mel-frequency cepstral coefficients (MFCC) as the acoustic features. We have compared different spectral features for automatic speaker identification on several databases, including subband analysis, mel- and Bark-warped cepstra, LPC-based features and formant frequencies, along with their delta features (Kinnunen *& al.* 2004a, 2004b). MFCC's have shown high accuracy in our experiments for both laboratory- and telephone-quality speech.

Speaker profiler uses 12 lowest MFCC coefficients, excluding the 0th coefficient, which is not a robust parameter as it depends on the intensity. The filterbank employed in the MFCC computation consists of 27 triangular filters equispaced on the mel frequency scale. A frame rate of 100 frames per second is used, with a 20 ms overlap between the adjacent frames (30 ms frame length, 10 ms frame shift).

Speaker matching is based on a vector quantization (VQ) approach (Kinnunen *& al.* 2004c). For each enrolled speaker, a codebook of size 64 is created using the K-means algorithm. During the matching, an unknown sample is scored against the stored codebooks, giving a dissimilarity value for each speaker. For easy interpretation, the scores are normalized into the interval [0,1] so that a larger score means better match. The program displays a ranked list of similarity values. Figure 1 shows a Winsprofiler screenshot. *Juhani* is speaking to a microphone (connected to PC-soundcard). Feature vectors computed from the speech are scored on-line against the database of 8 models, 7 were trained using sound files. Juhani was trained using the PC-soundcard.

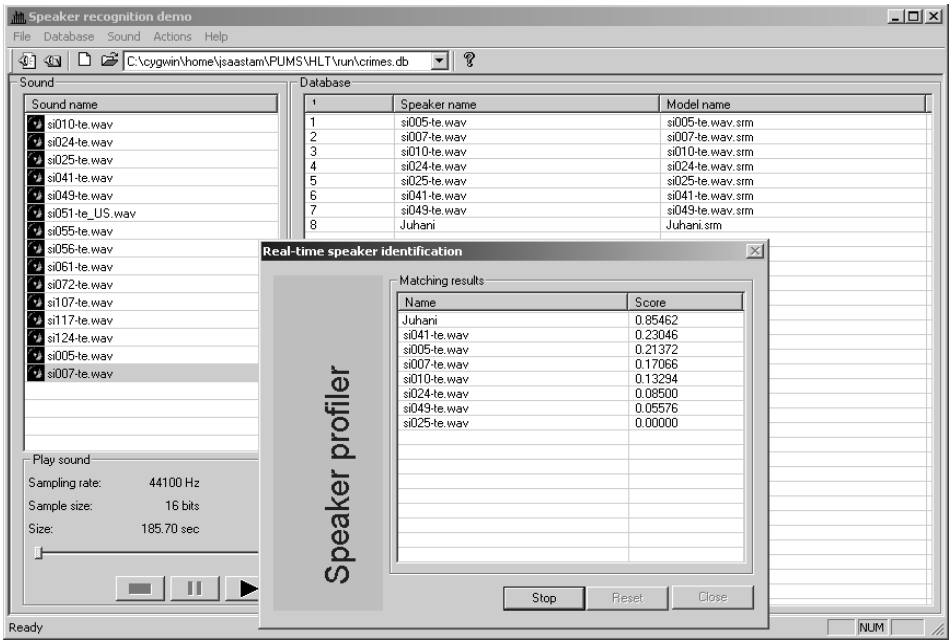

Figure 1. *Winsprofiler* real-time matching, *Juhani*'s speech is recorded and matched on-line against 8 database entries, including *Juhani*'s voice model

4. Test results with GSM data

In the first test situation spontaneous speech vs. reading was tested. Here the question was, does the program find the right person from the database when his/her two different, non-verbatim, speech samples are compared. First the database was created from the spontaneous speech samples and then, for reliability reasons, the database was also created from the text reading samples.

When the speaker database was constructed from the spontaneous speech samples, and the text reading samples were matched, there were 51 out of 107 tested samples ranked first (47.7 %), 68.2 % ranked among the first 3 samples, and 76.6 %

among the first 5 samples. The database contained 16 samples that were somewhat noisy or the voice of the speaker was either creaky or had very low pitch. After removing these 16 samples from the database, total of 91 samples formed a new database. The upper part of Table 2 shows the results of this test. When both noisy and nonnative speakers samples were removed, 53.2 % were ranked first, 73.4 % were ranked among the first 3 samples, and 81 % among the first 5 samples.

The speaker database was also formed from the text reading samples to check the reliability of the results. Total of 107 samples formed the database. When spontaneous speech samples of the same speakers were used as the reference, there were 25 samples out of 107 tested ranked first (23.4 %), 39.3 % was ranked among the first 3 samples, and 51.5 % among the first 5 samples, see Table 2 (lower part). When noisy and nonnative speakers' samples were removed, 62 % were ranked first. 76 % were ranked among the first 3 samples, and 87.3 % among the first 5 samples (see Table 2).

Table 2. Rankings of the correct speaker. The database is constructed from spontaneous speech samples and the testing material from text reading of the same speakers (upper part), and vice versa (lower part)

Read speech		Rank of the correct speaker		
	# Samples	1	1-3	1-5
All samples	107	47.7 %	68.2 %	76.6 %
Noisy samples removed	91	49.5 %	70.3 %	80.2 %
Noisy + non-native removed	79	53.2 %	73.4 %	81.0 %
Spontaneous		Rank of the correct speaker		
	# Samples	1	1-3	1-5
All samples	107	23.4 %	39.3 %	51.5 %
Noisy samples removed	91	57.1 %	73.6 %	83.5 %
Noisy + non-native removed	79	62.0 %	76.0 %	87.3 %

Poor quality of some of the recordings lowered the recognition score and the ratings. Noisy samples could affect many matching situations, where the recognition is not clear. Some tested speech samples were uttered in Finnish, but the speaker's mother tongue is Russian, Estonian, or Swedish (many Finns have Swedish as their mother tongue). This had some effect on the results. The reading and spontaneous speech samples of these speakers did not match as often as the others'.

5. Test results with real crime data

The second part of this study consists of real crime investigation cases. The known speech samples (suspect) form a speaker database. The unknown speech sample(s) recorded at the crime scene, were used as testing material. The test was done also vice versa: the unknown crime scene speech samples formed the speaker database and the known speech sample(s) from the suspects were used as testing material.

In the first case, 28 known speech samples from suspects formed the database, and 61 different phone calls from 13 different criminal cases were matched against the database. All the speech material in the criminal cases was recorded via GSM or land-line phones. From the matched 61 samples, 68.9 % were ranked first, 82 % among the first three, and 85.2 % among the first five. Three cases had either very short samples or the samples were noisy. These results are shown in Table 3 separately.

The testing with the forensic data was done twice. When the database consisted of 68 unknown criminal speech samples from 13 different crime investigation cases, and

the test material consisted of 25 known speech samples from suspects, the result was much worse. Only few samples of suspects and criminals matched. This situation is not typical. Databases consisting of unknown speakers are not used in forensics.

Table 3. Rankings of the correct speakers and the corresponding correct identification rates. The database is constructed from the 28 known samples (upper part) and from the 68 unknown samples (lower part)

Known samples		Rank of the correct speaker		
	# Samples	1	1-3	1-5
All samples	61	68.9 %	82.0 %	85.2 %
Noisy and short samples removed	58	72.4 %	86.2 %	89.7 %
Unknown samples		Rank of the correct speaker		
	# Samples	1	1-3	1-5
All samples	25	40.0 %	52.0 %	68.0 %
Noisy and short samples removed	22	45.5 %	59.0 %	72.7 %

6. Conclusion

There were some interesting findings in the spontaneous and text reading tests. The creaky voiced female speakers were often recognized as someone else. Likewise, the males with very low-pitched and/or creaky voice succeeded similarly. Some speakers in the database were ranked first many times when the samples from another speakers were matched against the database. The reason for this should be studied.

All the speech material in the criminal cases (test-2) was recorded via GSM or land-line phones and no extra background noise was found. This could be an error source for the study. Still the question arises, does the matching use more information from the background than from the speech signal itself.

Different results arise with the two opposite forensic tests, though the same samples are used in both. When the database is formed from known, usually long speech samples, the result is better than with the database consisting of unknown speech samples that are usually quite short. Is the sample duration the reason for this problem? The database sizes were different also, the recognition result was worse for the larger.

Currently the *Speaker Profiler* implements a baseline MFCC-VQ recognition. According to our experience, is quite accurate for matched conditions. However, more accuracy is desired in acoustically mismatched conditions, such as technical mismatches in recording. Further studies are therefore needed to detect the criticial conditions, and to employ methods for noise-, channel-, or score normalization.

The recognition performance of the Speaker Profiler is nowhere near perfect. However, it can be already as such used to create *shortlists,* i.e. pick substantially smaller sets of best ranked speaker candidates. The correct speaker is consistently found in the shortlist. This property already makes it a valuable tool for a crime investigator.

Acknowledgements

The Joint Project of Finnish Speech Technology was supported by TEKES, Civil Aviation Administration in Finland, The Finnish Defence Forces, National Bureau of Investigation in Finland, Accident Investigation Board Finland and Scando Oy.

References

Alexander, A., Botti, F., Dessimoz, D. And Drygajlo, A. 2004. The effect of mismatched recordings on human and automatic speaker recognition in forensic applications. *Forensic Science International* 146S (2004) S95-S99.

Botti, F., Alexander, A. And Drygajlo, A. 2004. On compensation of mismatched recording conditions in the Bayesian approach for forensic automatic speaker recognition. *Forensic Science International* 146S (2004) S101-S106.

Broeders, A.P.A. 1995. The role of automatic speaker recognition techniques in forensic investigation. *Proceedings of the XIIIth International Conference of Phonetic Sciences,* vol. 3, 154–161. Stockholm: KTH & Stockholm University.

Iivonen, A., Harinen, K., Keinänen, L., Kirjavainen, J., Meister, E. & Tuuri, L. 2003. Development of a multiparametric speaker profile for speaker recognition. *15th Int. Congress of Phonetic Sciences, Barcelona* 3-9 Aug, 2003, 695–698.

Kinnunen, T. (2004a). *Spectral features for automatic text-independent speaker recognition.* Licentiate's Thesis, Dept. of Computer Science, Univ. of Joensuu.

Kinnunen, T., Hautamäki, V., and Fränti, P. (2004b). "Fusion of spectral feature sets for accurate speaker identification", *Proc. 9th Int. Conference Speech and Computer* (*SPECOM'2004*), pp. 361-365, St. Petersburg, Russia, September 20-22, 2004.

Kinnunen, T., Karpov, E., and Fränti, P. (2004c). Real-time speaker identification and verification. Accepted for publication in *IEEE Transactions on Speech and Audio Processing*.

Künzel, H., Gonzáles-Rodríguez, J. 2003. Combining automatic and phonetic-acoustic speaker recognition techniques for forensic applications. *15th International Congress of Phonetic Sciences, Barcelona* 3-9 Aug, 2003, 1619–1622.

Niemi-Laitinen, T. 2001. Automatic speaker recognition – is it possible or not? In S. Ojala & J. Tuomainen (eds) *Papers from the 21st Meeting of Finnish Phoneticians* – Turku 4-5.1.2001. Publications of the Department of Finnish Language and General Linguistics, University of Turku 67: 71-80.

TUIJA NIEMI-LAITINEN is a researcher at the Crime Laboratory at the National Bureau of Investigation, Finland. Her research interests concern speech prosody and speaker recognition. E-mail: tuija.niemi@helsinki.fi or tuija.niemi-laitinen@krp.poliisi.fi

JUHANI SAASTAMOINEN is a project manager in the Department of Computer Science in the University of Joensuu, Finland. His current research interest is numerical methods in speech analysis. E-mail: juhani@cs.joensuu.fi

TOMI KINNUNEN is a doctoral student in the department of Computer Science in the University of Joensuu. His research topic is automatic speaker recognition. E-mail: tkinnu@cs.joensuu.fi

PASI FRÄNTI is a professor of Computer Science in the University of Joensuu, Finland. His primary research interests are in image compression, pattern recognition and data mining. E-mail: franti@cs.joensuu.fi