On Factors Affecting MFCC-Based Speaker Recognition Accuracy

Juhani Saastamoinen, Zdenek Fiedler, Tomi Kinnunen, Pasi Fränti

Department of Computer Science, University of Joensuu
Speech and Image Processing Unit
P.O.Box 111, 80101 Joensuu, Finland
{juhani,zfiedler,tkinnu,franti}@cs.joensuu.fi

Abstract

We evaluate the accuracy of an MFCC-based speaker recognition method. We analyse the recognition results using speech signal from everyday life environments. We study the mismatch effects of text-dependency, sample length, language, style of speaking, cheating, microphone, sample quality, and noise. The experiments on a self-collected corpus of 30 subjects indicate that any mismatch degrades recognition accuracy. The most dominating factors are noise, microphone, disguise, and degrading of the sample rate and quality. Speech-related factors and sample length are less critical.

1. Introduction

Accuracy of automatic speaker recognition is known to degrade severely when there is *acoustic mismatch* between the training and matching material [1, 2]. The mismatch can be due to the person himself (health, attitude), due to technical reasons (microphone, transmission channel), or due to the recording environment (additive noise, echo).

We have developed an automatic speaker recognition software system called *Sprofiler*, which consists of a portable C library with signal processing and pattern recognition engine [3]. The recognition is based on *vector quantization* (VQ) based speaker modeling and the features are *mel frequency cepstral coefficients* (MFCC). In the current implementation, we use closed-set speaker identification scenario for simplicity. The system is in pilot use at the Finnish National Bureau of Investigation Crime Laboratory, and we have reported preliminary experiments on the usefulness of the MFCC method in forensic casework [4]. Not surprisingly, we observed that performance is poor in the presence of mismatch.

In this paper, our motivation is to gain more understanding on factors affecting MFCC-based recognizer performance. Although the MFCC features are not necessarily the best choice, they are the most widely used feature set in text-independent speaker recognition. Our experiments with the mobile phone port of the Sprofiler [3] have revealed a large gap between corpus simulations and real conditions, even in noise-free environments. This discrepancy motivates us to find out the possible reasons.

We study the effect of

1. **Linguistic and data-related factors**:
   - Text dependency
   - Sample length
   - Language
2. **Speaker-dependent factors**:
   - Text reading vs. spontaneous speech
   - Disguise (deliberate cheating)
3. **Technical factors**:
   - Mismatched microphone types
   - Distance to microphone
   - Sampling rate
   - Additive noise

Regarding linguistic and data-related factors, we study whether the text content is important or not; in applications, text-independence is more convenient. The length of the sample is considered important, and we should confirm this. Regarding the language, we study if the speaker is better identified in his native language speech or in foreign language (English). We investigate if there is difference between the models which are trained in native and non-native speech.

From the speaker-related factors, we study whether the speech guidance (spontaneous vs. read) affects performance. In addition, we study disguise, i.e. speaker does not want to be recognized as himself and deliberately changes his voice.

From the technical factors, we study the effect of microphone mismatch, because this has been systematically reported in literature to be one of the main factors for degradation. We also study the effect of distance from the microphone. A close-talking microphone is expected to be more accurate, but less user-convenient. We also study the effect of sampling rate and additive background noise.

2. Test Setup

2.1. Recording Apparatus

Acer TravelMate 8000 series notebook was used in the recordings, with a 44.1 kHz and 16 bps sampling. All samples were stored as PCM-encoded WAV files. The recording volume was adjusted during the information chat with the speaker, but the speaking style variation changes the loudness in many samples. The notebook has a built-in sound card with a Realtek AC97 codec ALC202. Two different microphones were used in the recordings:

- **M1**: Integrated to a Plantronics headset .Audio$^{\mathrm{TM}}$80,
- **M2**: Built-in microphone of the notebook.

M1 is unidirectional, it has a noise-cancellation function, and the distance to mouth is fixed to 3–4 cm (headset). M2 is omnidirectional and the distance may vary between 50–70 cm.

Table 1: Nationalities of the speakers

Country		Count
CZE	Czech Republic	5
FIN	Finland	4
AUT	Austria	3
GER	Germany	3
POL	Poland	3
SPA	Spain	3
ROM	Romania	2
FRA	France	1
PRC	China	1
IND	India	1
INS	Indonesia	1
JPN	Japan	1
NEP	Nepal	1
NEL	Netherlands	1
Total number of speakers		30

2.2. Subjects and Tasks

We recorded speech from 30 individuals (16 M + 14 F) from 14 nationalities (Table 1). English is the common language for most speaking tasks but some tasks were spoken in the native language of each subject. Each speaker completed a set of compulsory tasks, each with a predefined set of sentences. They were recorded at 3–4 cm away from M1 and at 50–70 cm away from M2. We also study the effect of disguise as well as the spontaneous speech against text reading. Therefore, some subjects also completed additional speaking tasks where they were told to change their voice deliberately in order to be not recognized correctly, or speak spontaneously on an ordinary theme, such as the weather or personal feelings.

Our goal was to get the speaker familiar with the tasks, read the paper with the sentences, and answer the questions but not worry about the pronunciation or the translation of the sentences. The speaker must concentrate more on the speech itself instead of the fact that he or she is being recorded. For this reason, the spontaneous samples were recorded last, and the first 15 seconds of these samples are not used in the experiments.

2.3. Material

The speakers were asked to speak in English and in their native language. All speakers spoke the same sentences (Tables 6–7), but speakers had to translate the sentences in Table 7 to their native language themselves before speaking. Sentences were chosen with a particular interest in the occurrence of common English phonemes. The material is taken from [5, 6, 7]. Duration is less than 5 seconds for S03, S04, S06, and more than 15 seconds for the rest. Later we refer to "short" and "long" samples correspondingly. The spontaneous speech recordings are more than 90 seconds long.

Speaker models were always trained from speech material consisting of the sentences S01 (Table 6). We distinguish between text-dependent and text-independent utterances. All of the comprehensive recognition tests are based on sentences S02, S03, and S04, except for the text-dependent and language mismatch tests. Independent set of recordings of the sentences S01 were used for text-dependent recognition tests. In the language mismatch tests the recognition was based on the sentences S05 and S06 spoken in the native language of each subject.

2.4. Sample Preparation

After recording, we prepared the samples for the test runs. Each sample is trimmed by removing silence from both ends of the sample, signal is downsampled, and finally noise is added. We resampled and quantized the files using the SoX software to

- **A-quality**: 44.1 kHz, 16 bits,
- **B-quality**: 22.05 kHz, 16 bits,
- **C-quality**: 8 kHz, 8 bits.

We recorded samples of 5 different types of noise using the A-quality, 45 seconds each: babble, knock, rain, train, and ticks, as well as additional samples obtained from [8]. The babble noise simulates background talk. The knock is a repeated impulse every 0.740 s (knock on a wood desk). The rain noise is heavy rain on a window. The train noise is a sound of a train passing by, a repeated pattern on a noisy background. Ticks consists of random knocks and ticks on a wood desk with a 3 dB cut down rain noise in the background. Each sample has low- and high-volume versions with 6 dB intensity difference. The low ones are mixed with the M2 samples and the high ones with the M1 samples.

2.5. Parameter Setup

In all experiments, feature extraction parameters were fixed as follows. A 30 ms Hamming-windowed frames with 10 ms overlap are pre-emphasized. From the non-silent frames, FFT magnitude spectrum is smoothed using mel-scaled triangular-shaped 30 bandpass filters, and DCT is applied to the the log-compressed outputs. The lowest 16 coefficients are retained, and the zeroth coefficient is dropped because it depends on the intensity.

During the speaker model creation, VQ codebooks of size 256 are generated by the *generalized Lloyd algorithm* (GLA) [10]. Closed-set identification is performed by selecting the codebook yielding the smallest quantization distortion for the test vectors [11].

3. Results and discussion

3.1. Linguistic and data-related factors (D)

The linguistic and data-related factors which we studied are language mismatch between training and recognition, text-dependence, and the length of the sample. The effects of D-factors in recognition error rate are listed in Table 2, with varying training and recognition noise conditions.

3.2. Speaker-related factors (S)

Deliberate cheating is possible, error rates are 80–100 %. Error rates are similar when recognizing speakers from spontaneous speech, when the database is constructed from text reading. Recognition fails mostly, with or without noise mismatch.

3.3. Technical factors (T)

Noise has the strongest effect in recognition rate. It is evident from Tables 2–5. We looked closer into different noise types while using two different microphones, the effects are listed in Table 3. The effect of microphone is not significant, except for the mismatch of clean sample training samples and recognition from samples contaminated by impulsive noise. Other T-factor effects are described in Tables 4–5. The microphone distance

Table 2: Linguistic and data-related factor tests under two noise conditions. The average (AVG) and standard deviation (SD) of the recognition error rate (%) are computed for N recognition tests, data from both microphones M1 and M2 is used

	Same noise			Noise mismatch		
Factor	AVG	SD	N	AVG	SD	N
Text-dependent	0.74	2.10	18	63.9	26.6	90
Text-independent	3.64	4.90	54	66.1	25.8	270
Short sample	5.37	7.13	54	66.4	25.6	270
Long sample	4.88	6.18	54	69.6	25.2	270
Lang. mismatch	9.54	7.94	36	72.8	23.5	180

Table 3: Effect of various noise conditions with different microphones, N = 6 in all cases. The microphone effect is insignificant, except for the clean/knock noise mismatch

Noise type used in		Microphone M1		Microphone M2	
training	recognition	AVG	SD	AVG	SD
clean	clean	0	0	2.78	3.56
babble	babble	6.67	5.44	7.78	3.14
clean	knock	0	0	27.2	7.80
clean	babble	78.9	8.53	82.2	7.11

is missing, because we did not have enough samples to test its effect reliably.

3.4. Special tests

Besides the comprehensive testing, we also the tested certain particular features in *native training tests* and so-called *abracadabra tests.* We used the same stand-alone test framework as in [3]. It measures the exact effect of a single parameter change.

The idea of *native training test* comes up from the possible difference between databases constructed from native or foreign language speech. The question is whether the native speech perceives the voice model better than the foreign and therefore provides better recognition. The training data was S05 sentences recorded with M1 and sample quality B. Result of the 28 recognition tests: In noisy background, the recognition is better for non-native training data independent of the data type used during the matching.

The abracadabra test is focused on a proper distinguishing between text-dependent and text-independent recognition. Suppose that during database constructing, one user speaks ”abracadabra” and all the others use different words. What happens if all speakers say ”abracadabra” during matching? The situation was simulated in two tests using the stand-alone test framework:

Test 1: Train speaker 01 with sample ”M1, S02 recorded for matching”, all others with ”M1, S01, training. Match speaker 01 with ”M1, S01, training”, all others with ”M1, S02, matching”.

Test 2: Train similarly as in Test 1. Match all speakers with ”M1, S02, matching”.

Result of the 2 tests: recognition is text-independent.

3.5. Summary of the results

The descending order of significance of the factor effects is

1. noise (T),
2. different microphone (T),
3. disguise (S),
4. quality of the sample (T),
5. text reading contra ordinary speech (S),
6. sample length (D),
7. language (D),
8. text-dependency (D).

Table 4: Noise conditions and text-dependence (TD) tests under different microphone matching conditions

		Same microphone			Mic. mismatch		
Noise	TD	AVG	SD	N	AVG	SD	N
match	yes	0.74	2.10	18	84.4	7.86	18
match	no	3.64	4.90	54	85.0	8.95	54
mismatch	yes	63.9	26.6	90	86.8	13.2	90
mismatch	no	61.1	27.3	45	87.7	11.9	270

Table 5: Noise conditions and text-dependence (TD) tests under two different sample qualities

		Sample quality B			Sample quality C		
Noise	TD	AVG	SD	N	AVG	SD	N
match	yes	0.74	2.10	9	0.74	2.10	9
match	no	5.33	6.30	45	6.67	7.44	45
mismatch	yes	88.4	12.1	45	85.2	14.0	45
mismatch	no	90.6	9.00	225	86.0	12.4	225

Most results are influenced by many factors simultaneously. Computing factor specific effects would be misleading. A more detailed interpretation of each factor is described below.

- **noise (T)** — The background noise is the most significant factor for the recognition accuracy, which is high for the clean samples but deteriorates quickly for noisy samples. Only impulsive noise has no significant influence. Results are better without mismatch.
- **different microphone (T)** — Results are best without mismatch, the microphone quality itself is insignificant.
- **disguise (S)** — Deliberate cheating is possible, the recognition fails in most of the cases.
- **quality of the sample (T)** — Several cases have to be considered. Generally, the worst impact is caused by a background noise in connection with a microphone mismatch. For clean samples, the higher quality leads to the better results. However, B quality gives almost the same results as A. For noisy samples, the C quality produces occasionally better results, especially with M2. The A quality is unusable with noisy files. Not all possible combinations are mentioned, just a few examples
- **reading contra ordinary speech (S)** — The recognition with spontaneous speech fails for most tests.
- **sample length (D)** — The assumption that longer samples improve recognition, could not be verified. There is no significant difference to clean samples. In background noise the short samples provide better results but the difference is within the confidence level.

- **language (D)** — There is no advantage in native language speech. For the noisy samples, the non-native samples give better results.
- **text-dependency (D)** — It is confirmed that the recognition is text-independent for guided speech. The abracadabra test supports this. In a noisy background, difference is noticeable but is within the confidence level.

4. Conclusions

The most expected conclusion is that the training and recognition conditions should match. The most significant single changing factor, among the tested ones, is the noise. When the training and recognition data contain different types of noise, the error rate is above 75 % in most cases. When the noise conditions match, the error rate is systematically below 15 %.

Speech and language factors are less important than technical factors but deliberate cheating makes an exception: cheating is possible. The Speaker Profiler speaker recognition is text-independent. This conclusion is supported by both the conprehensive tests and the abracadabra test.

5. Acknowledgements

The research was carried out in the project *New Methods and Applications of Speech Processing* [7] and was supported by the Finnish Technology Agency. We thank all the volunteers who gave their voice for the sake of science.

6. References

[1] Lamel, L.F. and Gauvain, J.L., "Speaker Verification over the Telephone", *Speech Communication*, 31 (2000): pp. 141–154.

[2] Ortega-Garcia, J., González-Rodriguez, J., et al., "AHUMADA: A large speech corpus in Spanish for speaker identification and verification", *IEEE Intl. Conf. on Acoust. Speech and Signal Proc.*, (ICASSP-98), pp. 773–776, Seattle, May 1998.

[3] J. Saastamoinen, E. Karpov, V. Hautamäki, P. Fränti, "Accuracy of MFCC based speaker recognition in Series 60 device", *EURASIP Journal on Applied Signal Processing*, Accepted for publication.

[4] T. Niemi-Laitinen, J. Saastamoinen, T. Kinnunen, P. Fränti, "Applying MFCC-based speaker recognition to GSM and forensic data", *Proc. 2nd Baltic Conf. on Human Language Technologies (HLT'2005)*, pp. 317-322, Tallinn, Estonia, Apr 4–5, 2005.

[5] SAMPA, Dept. of Phon. & Ling., University College London, http://www.phon.ucl.ac.uk/home/sampa/home.htm, Sep 2004.

[6] EU Commission Human Language Technologies home page, http://www.hltcentral.org, Sep 2004.

[7] PUMS (New methods and applications of speech tech.), Dept. of Computer Science, Univ. of Joensuu, http://cs.joensuu.fi/pages/pums, Sep 2004.

[8] SPEECON project, http://www.speecon.com, Oct 2004.

[9] Rabiner, L. and Juang, B.-H., Fundamentals of Speech Recognition, Prentice Hall, 1993.

[10] Linde, Y., Buzo, A., and Gray, R. M., "An Algorithm for Vector Quantizer Design", *Trans. IEEE Commun.*, Vol. COM-28, pp. 84–95, 1980.

[11] Kinnunen, T., Karpov, E., and Fränti, P., "Real-time speaker identification and verification", *IEEE Trans. on Speech and Audio Processing*, Accepted for publication.

Appendix

Table 6: Sets of sentences spoken in English

S01	Is this a road to the Rocky Mountains?
	We all may hear a yellow lion roar.
	Every salt breeze comes from the sea.
	Shifting shelter is the shabby shelter.
	I have oiled the wheel with oily grease.
S02	Do not ask me to carry an oily rag like that.
	That mean of transport is better.
	Why do I owe you a letter?
	A birch on a sandy beach is swaying in the breeze.
	Do those cookies in your own oven.
	A black cat is rumbling on the roof.
S03	We all may hear a yellow lion roar.
	Shifting shelter is the shabby shelter.
S04	Should we chase?
	My mother is wandering around the rock.

Table 7: Sets of sentences spoken in Native language

S05	I am waiting here for 10 minutes.
	Have you understood me properly?
	I really do not want to read that book.
	What are you planning to do the next week?
	This is a job for you.
	Pardon, could you show me the way again, please?
S06	Open the door, please.
	Could I enter?
	I am here to introduce myself.