Intelligibility Assessment of the De-Identified Speech Obtained Using Phoneme Recognition and Speech Synthesis Systems

Tadej Justin, France Mihelič, and Simon Dobrišek

Faculty of Electrical Engineering, University of Ljubljana,
1000 Ljubljana, Tržaška 25, Slovenia
{tadej.justin, france.mihelic,simon.dobrisek}@fe.uni-lj.si
http://luks.fe.uni-lj.si

Abstract. The paper presents and evaluates a speaker de-identification technique using speech recognition and two speech synthesis techniques. The phoneme recognition system is built using HMM-based acoustical models of context-dependent diphone speech units, and two different speech synthesis systems (diphone TD-PSOLA-based and HMM-based) are employed for re-synthesizing the recognized sequences of speech units. Since the acoustical models of the two speech synthesis systems are assumed to be completely independent of the input speaker's voice, the highest level of input speaker de-identification is ensured. The proposed de-identification system is considered to be language dependent, but is, however, vocabulary and speaker independent since it is based mainly on acoustical modelling of the selected diphone speech units. Due to the relatively simple computing methods, the whole de-identification procedure runs in real-time.

The speech outputs are compared and assessed by testing the intelligibility of the re-synthesized speech from different points of view. The assessment results show interesting variabilities of the evaluators' transcriptions depending on the input speaker, the synthesis method applied and the evaluators capabilities. But in spite of the relatively high phoneme recognition error rate (approx. 19%), the re-synthesized speech is in many cases still fully intelligible.

Keywords: Voice de-identification, phoneme recognition, speech synthesis, diphone speech units, HMM modelling, intelligibility evaluation.

1 Introduction

De-identification can be defined as the process of concealing the identities of individuals captured in a given set of data (e.g., video, audio or text) for the purpose of protecting their privacy [1]. De-identification techniques should not only be capable of preventing humans from recognizing subjects in the multi-media content, they should also be able to conceal identities from automated recognition systems, such as face recognition [2], speaker verification and others.

Many different techniques to obtain the speaker's de-identified voice have already been reported. We can divide them into techniques using a voice-degradation approach,

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 529–536, 2014.

a voice-conversion approach or a voice-morphing approach [3,4,5]. These techniques can be used in real applications, but each with their own limitations, which have to be considered. The voice obtained with the voice degradation technique often runs as an on-line process, but the resulting output speech is more or less unnatural. The aim in the voice conversion approach is to assess the mapping transformation from the input speaker to the target speaker. For such an operation the access to the input speaker and the target speaker audio recordings and/or acoustic models must to be provided in advance. With the goal to obtain an on-line speaker de-identification system, we propose and investigate the intelligibility of a language-dependent method, which is lexical independent, but still uses the possibility of a speech synthesis system. Despite the fact that it is not possible to obtain robust automatic word recognition only from a phoneme-recognition system (a phonetic typewriter) the evaluation of automatically recognised phonemes shows that errors made by the recognition system are often realized as substitutions between phonetically similar phonemes. By listening to an utterance with such substitution errors, the listener with a suitable linguistic knowledge can still recognise most of the uttered words and understand the meaning. If we assume that the speech synthesis system uses its own target speaker's voice, which is different from the input speaker, then the output voice can be seen as a de-identified input voice.

The system is described more precisely in the next section. Since the intelligibility of the speech generated using the presented approach is questionable, our experiment was focused on a subjective assessment of the intelligibility of the de-identified speech. We describe and comment on the evaluation procedure in Sections 3 and 4. We summarize and further discuss our experimental results in the conclusion, where we also propose further steps for improvements.

2 System Description

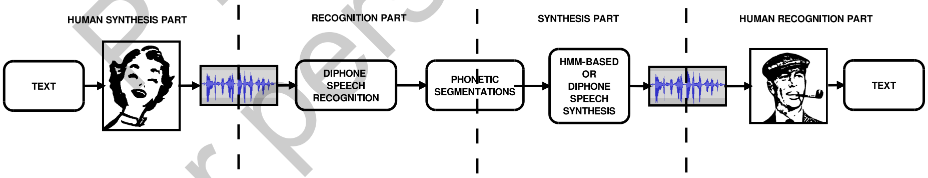

Fig. 1. Speaker de-identification system evaluation scheme

The experiments were conducted with two different configuration set-ups (Figure 1). In each set-up we used our own implementation of the Slovene phoneme-recognition system [6]. The system was developed with the use of Hidden Markov Models (HMMs) and bigram phoneme language modelling. In our implementation, the basic acoustic units were presented as context-dependent diphones, called bi-diphones. These diphone units are defined as speech segments containing mostly a transition between the two neighbouring phonemes and were in the beginning introduced in the field of speech synthesis [7]. Bi-diphone recognition provides better results in comparison with the classic triphone HMM-based recognition on the same evaluation test corpora that was used for our present evaluation study [8].

The Slovene speech synthesis systems used in the experimental set-ups have also been developed in our institution. Each of the two speech synthesis systems was trained with a different speech database; therefore, the synthesised voice from one or other system has different target-speaker characteristics. In the first set-up the de-identified voice was obtained with the use of a diphone speech synthesis system [9]. In parallel, the second set-up represents the use of a HMM-based speech synthesis system [10] to achieve the same goal. A second system was developed with the use of the HTS toolkit, version 2.2 [11], where contextual quin-phones were used for the base units. It is well known and also proved in previous works, such as [12], that the synthetic speech synthesised with the HMM-based approach is more natural in comparison with the speech obtained with the diphone synthesis system. Nevertheless, the comparison of intelligibility of such speech systems still remains an open issue, especially in our case where diphone speech units were also the base units in our recognition system.

The output of a bi-diphone recognition system represents a string of Slovene allophones with their associated estimations of duration and provides the input to the speech synthesis systems. Pitch estimation and loudness can also be obtained from the recognizer, but since the output speech signal has to present the de-identified voice, this was discarded from subsequent processing. Speech synthesis systems could also not take into account any additional prosodic modelling based on the word units, which are usually obtained in text-to-speech systems.

3 System Evaluation Setup

The proposed system was tested on the Slovene speech database GOPOLIS [13], which contains the speech signals (read speech) and their transcriptions from 50 (25 male and 25 female) speakers. The word-recognition system using this database was primarily developed in our previous work [14], with the aim being to build an automatic speech dialogue system for querying flight information. Using the standard protocol, the training set contained recordings from first 18 male and 18 female speakers and a test set of the remaining 7 male and 7 female speakers' speech recordings. The training part of the database was used to train our bi-diphone speech recognition system.

For the voice de-identification system's evaluation we randomly picked 28 test sentences from the test set with the following limitations:

- Only two sentences from the same speaker.
- Each sentence had to be between 5 and 8 words long.

With such limitations we ensured that, on one hand, the synthesised sentences are not too short and, in addition, are not too simple to understand. And on the other hand, the sentences were not allowed to be too long and consequently easy to forget, since the evaluator's task was to remember and transcribe the recognised words from the artificial speech utterance.

Using both limitations we appropriately distributed the test sentences between all the different test speakers. The final evaluation set consisted of 2 · (7 different male) and

$2 \cdot$ (7 different female) input sentences, resulting in 56 (28 diphone synthesis, 28 HMM synthesis) different synthesised utterances.

All the evaluation tests were made with our own web-based evaluation system. Even though the application can be accessed from every computer with an internet connection and an Adobe–Flash–enabled web browser, the evaluation process took place in a controlled environment at the Faculty of Electrical Engineering the University of Ljubljana. During the evaluation process all the evaluators wore headphones. The evaluation test was successfully completed by 26 evaluators. All the evaluators were 3^{rd}–year university program students from the University of Ljubljana, Faculty of Electrical Engineering. The evaluators had a limited, basic knowledge of speech technologies.

Before the evaluation task was started a brief explanation of the developed system and the evaluation process was introduced to the evaluators. Between the introduction also the semantic-domain (flight-service information queries) of the evaluation utterances was explained. Each evaluator transcribed 7 sentences synthesised with the diphone speech synthesis system and 7 sentences synthesised with the HMM–based synthesis system. The evaluators were divided into two groups. The first group with 13 evaluators evaluated the first randomly picked sentence from each speaker in the test dataset. The second group evaluated the sentences that were not evaluated by the first group of evaluators. With such an evaluation set-up we ensured that each evaluator listened to each sentence only once and also that all the evaluator's transcriptions belong to different input speakers. With such an evaluation process we obtained $(13 + 13) \cdot 14 = 364$ transcriptions.

The evaluated system for speaker de-identification can produce two kinds of errors. The first one is related to the bi-diphone speech recognition system, and it can be measured as the Phoneme Error Rate (PER). The second one can be presented as the output system error. This error results as a combination of influences from errors made by the speech recognition system, the performance of the synthesis system and the evaluator's capabilities. It can be measured from the analysis of the evaluator's transcriptions in relation to the reference sentence's transcriptions as the Word Error Rate (WER).

The error rate (ER) is defined from the accuracy (A)

$$ER = 1 - A \quad , \tag{1}$$

where the accuracy is determined as in [15] from the number of correctly recognized units (N_{cor}), the number of insertion errors (I) and the number of reference units (N_{ref})

$$A = \frac{N_{cor} - I}{N_{ref}} \quad .$$

4 Experimental Results

The evaluation analysis results can provide some interesting conclusions about the human ability to recognize words based on the acoustic representation of phoneme

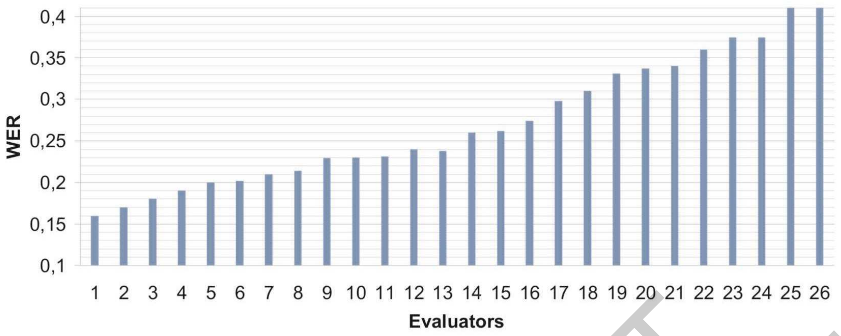

Fig. 2. Evaluators average word error rate (WER) for all the listening tests

recognition. Such a process of human perception is comparable to the process used in automatic systems for speech recognition, where we commonly realize it with the help of a phonetic word lexicon, sentence syntax and semantic language modelling.

Figure 2 presents the variance of 26 evaluators transcription WER capabilities. We can conclude that the results are strongly dependent on the the evaluator's identity. A detailed analysis shows that there are many transcriptions from the same evaluator's identity that are fully recognized (WER = 0) or with a very low WER, although - in terms of PER - not even one input sentence was completely correctly recognized by the bi-diphone speech recognition system. We can also observe this phenomenon in Figure 3.

Figure 3 also shows the trends of the average transcriptions WER for input sentences depending on the recognition PER. As expected we can observe positive linear regressions. From there we can also notice the difference between the average WER depending on the speech synthesis system. In the present voice de-indentification system the diphone speech synthesis system is more intelligible than the HMM-based speech synthesis system. Significant differences in the WER can also be confirmed from Table 1, displaying the average WER for different types of speech synthesis.

Table 1. Average word error rate (WER) for different types of speech synthesis

Number of listening tests	WER HMM	WER Dif
182	0.33	0.21

If we compare the obtained WER results with the WER from the words-recognition system presented in our previous work [14] describing a spoken dialogue system for air-flight queries, where the same acoustic models were used, we obtained - at first sight - a surprising paradox. The average WER 21 % obtained with the diphone speech synthesis

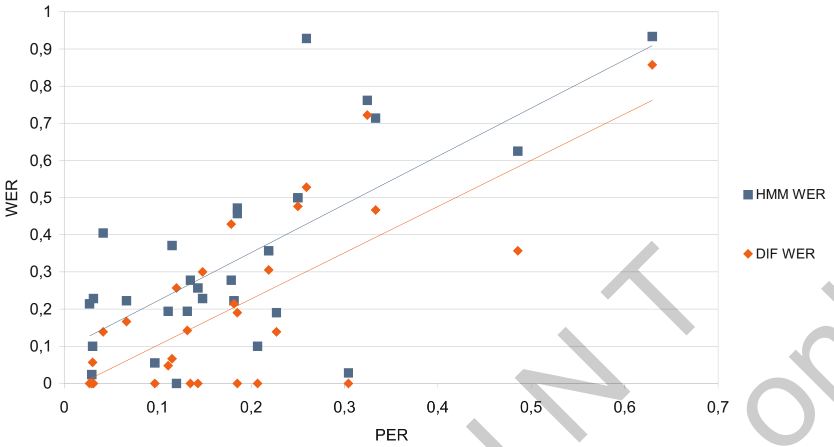

Fig. 3. Average word error rate WER per utterance for HMM synthesis and Diphone synthesis depending on the recognition phoneme error rate (PER)

Table 2. Correlations between (WER) and different types of phoneme recognition errors: deletions (D), substitutions (S), insertions (I) and phoneme error rate PER

	D	S	I	D + S + I	PER
WER HMM	0.71	0.40	0.34	0.71	0.69
WER Dif	0.58	0.53	0.40	0.76	0.73

is significantly higher than the WER 8% obtained from the dialogue system, thus the machine speech recognition system outperforms the human recognition abilities? In this case this apparent phenomenon can be explained by the fact that the word-recognition system used a relatively small word lexicon (829 words) and a syntax model with a very low perplexity (5.7) [13].

Based on the evaluation results we can roughly estimate the value of the PER from input sentence where the evaluators could still understand the de-identified synthesised utterance. Certainly, we do not need a 0% PER. In fact, in some occurrences the evaluators achieved the correct transcription (WER = 0), although the recognition PER was near to 50%. However the WER and PER are strongly correlated. If we further investigate the phoneme recognition errors' influences on the WER (Table 2), we can speculate that the phoneme deletions error have the major influence on de-identified voice intelligibility, especially for the HMM based synthesis (0.71 estimated correlation) and insertions are the least critical. It also seems that the error-measure defined as $D + S + I$ is more descriptive in our case than the usually obtained ER

Table 3. Average evaluators word error rate (WER) for different types of speech synthesis and average phoneme error rate (PER) for all test utterances, depending on the speaker's gender

gender	WER HMM	WER Dif	PER
female	0.44	0.29	0.23
male	0.23	0.13	0.14

measure, defined in (1). On the other hand, the PER and consequently the WER results are most likely also dependent on input the speaker's voice. For instance, we can see (Table 3) that the PER and, consequently, the WER are speaker–gender dependent.

5 Conclusion

In this article we propose an approach for a speaker-de-identification system and its evaluation with the use of subjective listening tests for intelligibility measuring. The evaluation results showed major differences between the transcriptions of the individual evaluators. Such results are not desirable, but also not critical for the evaluation of our proposed system. From the results it is evident that every result obtained from the subjective evaluations is strongly related to the evaluator's motivation in participating in the evaluation process. On the other hand it should be noted that in some cases due to a low recognition phoneme error rate, some sentences were not understandable to all the evaluators at all. This fact suggests the questionable usage of such a system in real applications.

Future improvements can be expected with the use of speaker-adaptation techniques in a speech recognition system, with the inclusion of additional application-dependent word spotting or even the replacement of the phoneme speech recognition system with a word recognition system. A possible solution to achieve better results could also be the use of pre-defined user training. Since it is an online system, the instant system de-indentified voice output could be used for training as an acoustic feedback.

Acknowledgments. The work presented in this paper was supported in parts by the national research program P2-0250(C) Metrology and Biometric Systems, and the European Union's Seventh Framework Programme (FP7-SEC-2011.20.6) under grant agreement number 285582 (RESPECT). The authors additionally appreciate the support of COST Actions IC1106 and IC1206.

References

1. Ribarić, S., et al.: De-identification for privacy protection in mutlimedia content. COST Action MOU (2013)
2. Poh, N., Štruc, V., Pavešić, N., et al.: An evaluation of video-to-video face verification. IEEE Transactions on Information Forensics and Security 5(4), 781–801 (2010)
3. Stylianou, Y.: Voice Transformation: A survey. In: ICASSP 1999, pp. 3585–3588 (1999) ISSN 1520-6149

4. Pfitzinger, H.R.: Unsupervised Speech Morphing between Utterances of any Speakers. In: Cassidy, S., Cox, F., Mannell, R., Palethorpe, S. (eds.) Proceedings of the 10th Australian International Conference on Speech Science & Technology, pp. 545–550 (2004)
5. Qin, J., Toth, A.R., Schultz, T., Black, A.W.: Speaker de-identification via voice transformation. In: IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2009, pp. 529–533 (2009) ISBN 978-1-4244-5478-5
6. Dobrišek, S., Mihelič, F., Pavešić, N.: Acoustical modelling of phone transitions: biphones and diphones - what are the differences? In: Olaszy, G., Nemeth, G., Erdohegyi, K. (eds.) Proceedings of Eurospeech 1999, vol. 3, pp. 1307–1310 (1999)
7. O'Shaughnessy, D., Barbeau, L., Bernardi, D., Archambault, D.: Diphone speech synthesis. Speech Communication 7(1), 55–65 (1988)
8. Dobrišek, S.: Analysis and Recognition of Phones in Speech Signals, PhD Thesis, University of Ljubljana (2001)
9. Žganec Gros, J., Pavešić, N., Mihelič, F.: Text-to-Speech synthesis: A complete system for the Slovenian language, vol. 5(1), pp. 11–19. CIT (1997) ISSN 1330-1136.
10. Pobar, M., Justin, T., Žibert, J., Mihelič, F., Ipšić, I.: A Comparison of Two Approaches to Bilingual HMM-Based Speech Synthesis. In: Habernal, I., Matousek, V. (eds.) TSD 2013. LNCS (LNAI), vol. 8082, pp. 44–51. Springer, Heidelberg (2013)
11. Zen, H., Nose, T., Yamagishi, J., et al.: The hmm-based speech synthesis system (hts) version 2.0. In: Proc. of Sixth ISCA Workshop on Speech Synthesis, pp. 294–299 (2007)
12. Vesnicer, B., Mihelič, F.: Evaluation of the Slovenian HMM-based speech synthesis system. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2004. LNCS (LNAI), vol. 3206, pp. 513–520. Springer, Heidelberg (2004)
13. Mihelič, F., Žganec Gros, J., Dobrišek, S., Žibert, J., Pavešić, N.: Spoken language resources at LUKS of the University of Ljubljana. Int. J. Speech Technol. 6(3), 221–232 (2003)
14. Ipšić, I., Mihelič, F., Dobrišek, S., Gros, J., Pavešić, N.: A Slovenian spoken dialog system for air flight inquiries. In: Olaszy, G., Nemeth, G., Erdohegyi, K. (eds.) Proceedings of Eurospeech 1999, vol. 6, pp. 2659–2662 (1999)
15. Young, S.J., Evermann, G., Gales, M.J.F., et al.: The HTK Book, version 3.4.1. Cambridge University Engineering Department, Cambridge (2009)