Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/FAIA200601

Using Privacy-Transformed Speech in the Automatic Speech Recognition Acoustic Model Training

Askars SALIMBAJEVS [1]
Faculty of Computing, University of Latvia, Latvia

Abstract. Automatic Speech Recognition (ASR) requires huge amounts of real user speech data to reach state-of-the-art performance. However, speech data conveys sensitive speaker attributes like identity that can be inferred and exploited for malicious purposes. Therefore, there is an interest in the collection of anonymized speech data that is processed by some voice conversion method. In this paper, we evaluate one of the voice conversion methods on Latvian speech data and also investigate if privacy-transformed data can be used to improve ASR acoustic models. Results show the effectiveness of voice conversion against state-of-the-art speaker verification models on Latvian speech and the effectiveness of using privacy-transformed data in ASR training.

Keywords. Automatic speech recognition, voice conversion, privacy, anonymization, evaluation, automatic speaker verification

1. Introduction

Voice-operated technologies and tools have multiplied in recent years; voice is rapidly replacing touch or text as the main means of interaction with modern devices.

These technologies require huge amounts of speech data to reach state-of-the-art performance. The standard today is to store the voices of end users in the cloud and label them manually. There are few guarantees (if any) regarding how data stored in the cloud is used and will be used in the future by cloud service providers. This approach raises critical privacy concerns and has led to market and data concentration in the hands of big corporations. Dramatic improvements in speech synthesis [1], voice cloning [2] and speaker recognition [3] pose severe privacy and security threats to the users.

This resulted in a growth of interest on new voice privacy-preserving transformations and voice privacy evaluations [4,5,6,7]. Recently, the VoicePrivacy initiative was started to spearhead the effort to develop privacy preservation solutions for speech technology and create a new community. [8].

The advancement of privacy-preserving methods enables the collection of anonymized speech data and raises at least two questions:

[1]Corresponding Author: Askars Salimbajevs; Tilde, Vienibas gatve 75a, Riga, Latvia, LV-1004; E-mail: askars.salimbajevs@tilde.lv.

1. Do these methods work for smaller and less-researched languages like Latvian?
2. Can privacy-transformed speech data be used to improve automatic speech recognition (ASR) acoustic models?

Therefore, first, we investigate the applicability of one of the voice anonymization methods (VoiceMask, [7]) to the Latvian language and evaluate the performance of Automatic Speaker Verification (ASV) on original and privacy-transformed speech.

Next, we train several ASR acoustic models on speech data processed by VoiceMask method. It is common among researchers to evaluate the intelligibility of privacy-transformed speech by calculating the word error rate (WER) of the ASR system. However, to the best of our knowledge, there is no research on the training of ASR acoustic models on privacy-transformed speech.

The question on the applicability of the language independent voice anonymization method for Latvian speech data might seem naive, however, we think that it is important to perform such validation as it shows that ASR acoustic models will be trained on a substantially different type of data with much of the speaker personality removed.

2. Evaluation Setup

2.1. Privacy-Preserving Voice Transformation

For the voice anonymization, we use VoiceMask voice conversion technique, which is proposed by Qian et al.[7,9]. After using standard signal processing methods to compute spectral envelope, pitch, and aperiodicity features, VoiceMask modifies the spectral envelope through frequency warping. To provide privacy, this method is based on the composition of a quadratic function and a bilinear function using two different parameters. The inverse of this transformation is much more difficult to compute, and, therefore, more resistant to attacks.

2.2. Automatic Speaker Verification

Automatic Speaker Verification (ASV) is the authentication of individuals by doing voice analysis on speech utterances. ASV has two phases: enrollment and verification. During enrollment, the speaker's voice is recorded and typically a number of features are extracted to form a voice print, template, or model. In the verification phase, a speech sample or "utterance" is compared against a previously created voice print to verify the identity of the speaker.

In this paper, x-vectors[3] are used as speaker voice-prints and PLDA[10] for x-vector classification. Model training and evaluation is done using the Kaldi toolkit [11].

In experiments with speaker verification, we use the 100h Latvian Speech Recognition corpus [13]. From this corpus, we select 50 speakers (28 - male, 22 - female) which are used to create a test set.

The test set is split into enrollment data and trials. Enrollment data consists of approximately 60 seconds of audio per each speaker recorded in different conditions. The remaining audio recordings of these 50 speakers are used for trials if they are recorded in conditions different from the enrollment. As a result, the total number of utterances

in the trial set is about 7,400, which results in 164,068 trials (trials are only between speakers of the same gender).

Audio recordings from other speakers in the corpus (approximately 1,700 speakers) are combined into a training set for the Latvian x-vector model. This training set is then augmented from 80h to about 375h by adding background noises and speed perturbations.

We evaluate ASV performance on the Latvian speech in the following settings:

- English x-vector and PLDA models trained on VoxCeleb2 dataset[12];
- X-vector and PLDA models trained on Latvian data only.

In each setting, we compare the performance of ASV system on original and privacy-transformed trial speech recordings using only original enrollment data.

2.3. Automatic Speech Recognition

The two crucial parts in the standard speech recogniser architecture are the acoustic model, which encodes pronunciation information, and the language model, which encodes grammar information. The open-source Kaldi toolkit [11] is used to train and evaluate Latvian ASR models.

For training of the ASR acoustic models, we use two speech corpora:

- The 100h Latvian Speech Recognition corpus [13] as a baseline training dataset.
- The Latvian Parliament Speech corpus [14] as additional data that is appended to ASR training dataset. A subset of 100 hours was taken to make the total length of both corpora comparable.

We train end-to-end Factorized Time Delay Neural Network (TDNN-F) [15] acoustic models with Lattice-Free Maximum Mutual Information (LF-MMI) [16] in a flat-start manner [17]. The model architecture and hyper-parameters are copied from the recipe for the Wall Street Journal (WSJ) dataset [18] which has a similar size (80 hours). Because Latvian has highly phonemic orthography, word pronunciation is modelled by treating each grapheme as a separate phoneme.

For language modelling, we employ a sub-word 4-gram language model which is trained on a 40 M sentence text corpus collected from Latvian web news portals. The model has a sub-word unit vocabulary generated using the Byte-Pair Encoding (BPE) method. N-grams are pruned to about 110 MB so that the decoding process can fit in 2 GB of RAM. Correct sub-word unit combination is ensured by a modified decoding graph [19].

Because the speech recordings which are appended to the baseline training dataset come from particular domain (Latvian Parliament session recording), testing was performed on two evaluation datasets: (1) an in-domain Saeima test set and (2) an out-domain test set of queries and short messages.

The in-domain Saeima test set contains 439 utterances (1 hour) from recordings of debates in the Parliament of Latvia from 2014 to 2016, containing contributions from about 300 different speakers. The recording time period does not overlap with the previously mentioned Latvian Parliament Speech corpus, as to guarantee that all utterances in the training and test sets are distinct (some overlap in speakers is still possible).

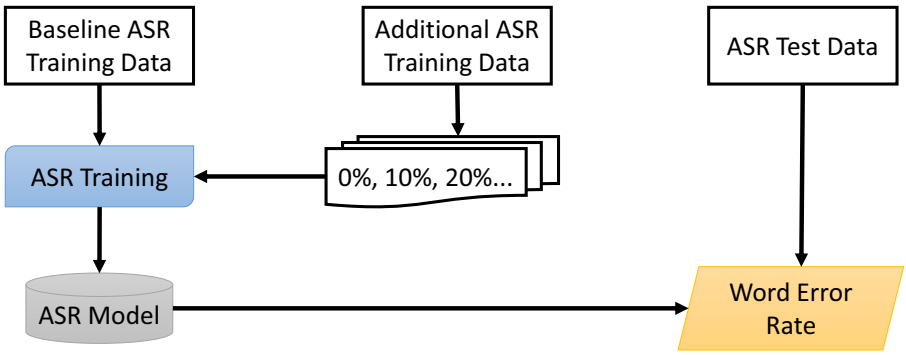

Figure 1. Baseline experimental setup of the speech recognition evaluation

To evaluate the effect of additional training data on the ASR performance in other domain, we use 1,159 utterances from real world data collected by production-level real-time Latvian ASR service.

2.4. Speech Recognition Using Anonymized Training Data

The experiment we devised compares two setups - one with and one without privacy-preserving speech transformation - with respect to the inclusion of additional training data on top of an existing baseline data set.

In the first setup, we add original, untransformed speech data to the baseline dataset by increments of 10 hours. We train a new ASR model for each increment and calculate the ASR performance on a test corpus (see Figure 1).

The second setup is similar to the first one. The only difference is that the additional training data is privacy-transformed prior to adding it to the baseline data (see Figure 2).

In both experiments, we add the respective portions of additional speech training data to the baseline dataset in the same order, i.e. ASR models are trained on exactly the same speech recordings in either setup. Also, we use untransformed test corpus in evaluation, because for privacy-preservation, we plan to operate ASR models on-device and adapt to the voice of the user. Voice anonymization will make such adaptation impossible.

3. Results

3.1. Speaker Verification

The speaker verification evaluation results presented in Table 1 show that VoiceMask voice transformation is effective on Latvian speech and can conceal the user identity (EER of ASV systems increased more than 3 times). This result corresponds to findings on the English data [4].

We can also observe that the state-of-the-art English x-vector model trained VoxCeleb2 corpus is language-independent and performs better than model trained on the Latvian data. We believe this is due to the fact that VoxCeleb2 corpus is more than 20 times larger than corpus used to train Latvian models.

Table 1. Speaker recognition evaluation on untransformed and privacy-transformed speech

Training data	Equal error rate %	
	Original speech	Transformed speech
VoxCeleb2	10.4	32.6
Latvian	11.8	32.6

3.2. Speech Recognition

First, ASR quality evaluation of models trained using different amounts of additional data was performed in-domain Saeima test set. The results presented in Table 2 indicate that both types of additional data improve the WER and the difference between adding transformed and untransformed data is small (5 % relative between best results of both methods).

Next, the evaluation on the second test set was performed to check if additional data can improve ASR performance on out-of-domain data (see Table 3). The results are quite noisy which may be attributed to a mismatch between the domain of the original training set and the additional data. Still, it is possible to make three main observations:

- Additional data improves speech recognition quality;
- Adding untransformed data helps to achieve better WER;
- The difference between adding untransformed and privacy-transformed data is small (2 % relative between best results of both methods).

There is a noticeable WER improvement after adding the first 10 hours of privacy-transformed data which seems suspicious. Interestingly, adding the same 10h of untransformed data only has a similar effect when evaluating on in-domain test data, but not on out-of-domain data. As an additional experiment, we decided to take the last 10h of additional data instead of first 10h and to retrain the system. This time the WER improvement was smaller and fitted with other results. Therefore, we believe this result

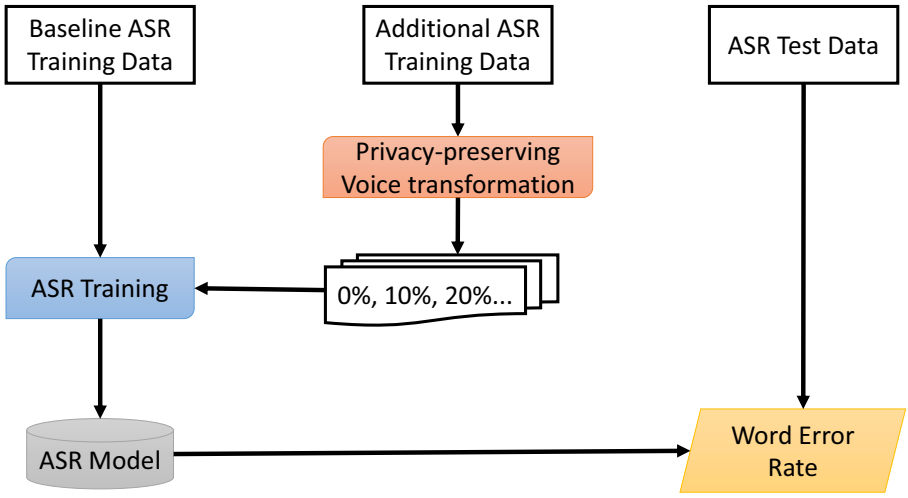

Figure 2. Experimental setup for the evaluation of the ASR trained using privacy-transformed data.

Table 2. Comparison of in-domain WER using untransformed and privacy-transformed speech in ASR training

New data, h	Word error rate %	
	Original speech	Transformed speech
0	13.6	13.6
10	12.8	12.7
20	12.5	13.3
30	12.5	13.1
40	12.4	13.1
50	11.9	13.0
60	11.9	13.0
70	12.1	12.6
80	12.2	12.7
90	11.8	12.4
100	11.8	12.8

Table 3. Comparison of out-domain WER using untransformed and privacy-transformed speech in ASR training

New data, h	Word error rate %	
	Original speech	Transformed speech
0	28.3	28.3
10	27.7	26.9
20	27.4	27.6
30	27.5	27.5
40	26.7	26.5
50	25.9	27.2
60	26.5	26.9
70	27.5	27.5
80	26.5	26.7
90	26.5	27.5
100	26.4	27.3

can be explained by irregularities in the additional data, some subsets of which are more beneficial than others.

4. Conclusions

In this paper, we investigated the use of VoiceMask voice anonymization method to protect the privacy of Latvian speakers by concealing their identity and also feasibility of using such transformed recordings in the ASR acoustic model training.

To the best of our knowledge this is a first evaluation of this kind. During the preparation of the paper, much better methods were created within Voice Privacy Challenge [8]. These will have to be evaluated in the similar way. Still, even at this early stage, the evaluation has now given us important insights.

Speaker verification experiments showed that VoiceMask method works for Latvian speech and can provide reasonable protection against attacks without knowledge of the anonymization method. This result also show that the privacy-transformed data is substantially different from the original and presents a new challenge to ASR acoustic model training.

We found that using such privacy-transformed in-domain data for acoustic model training resulted in a clear benefit for recognition quality. With a test set from another domain, the benefits of adding more training data suffered from noise artifacts. However, an improvement in WER still can be observed. While in both cases the benefit is smaller than when using the original speech data, we believe that this result proves that privacy-transformed speech data can be used to improve ASR acoustic models, therefore, allowing to collect the speech data from end users for training while preserving some privacy.

5. Acknowledgements

The work presented in this paper has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 825081 and under the name COMPRISE (Cost-effective, Multilingual, Privacy-driven voice-enabled Services).

References

[1] Székely É, Henter GE, Beskow J, Gustafson J. Spontaneous conversational speech synthesis from found data. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2019.

[2] Vestman V, Kinnunen T, González Hautamäki R, Sahidullah M. Voice Mimicry Attacks Assisted by Automatic Speaker Verification. Comput Speech Lang. 2020;

[3] Snyder D, Garcia-Romero D, Sell G, Povey D, Khudanpur S. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. 2018.

[4] Srivastava BML, Vauquier N, Sahidullah M, Bellet A, Tommasi M, Vincent E. Evaluating Voice Conversion-based Privacy Protection against Informed Attackers. In: ICASSP 2020 - 45th International Conference on Acoustics, Speech, and Signal Processing [Internet]. Barcelona, Spain; 2020.

[5] Lal Srivastava BM, Bellet A, Tommasi M, Vincent E. Privacy-preserving adversarial representation learning in ASR: Reality or illusion? In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2019.

[6] Ribaric S, Ariyaeeinia A, Pavesic N. De-identification for privacy protection in multimedia content: A survey. Signal Process Image Commun. 2016;

[7] Qian J, Du H, Hou J, Chen L, Jung T, Li XY. Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity. In: SenSys 2018 - Proceedings of the 16th Conference on Embedded Networked Sensor Systems. 2018.

[8] Tomashenko N, Srivastava BML, Wang X, Vincent E, Nautsch A, Yamagishi J, et al. The VoicePrivacy 2020 Challenge Evaluation Plan.

[9] Qian J, Du H, Hou J, Chen L, Jung T, Li X-Y, et al. VoiceMask: Anonymize and Sanitize Voice Input on Mobile Devices. 2017.

[10] Ioffe S. Probabilistic linear discriminant analysis. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2006.

[11] Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, et al. The Kaldi Speech Recognition Toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society; 2011.

[12] Chung JS, Nagrani A, Zisserman A. VoxceleB2: Deep speaker recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2018.

[13] Pinnis M, Auziņa I, Goba K. Designing the Latvian Speech Recognition Corpus. In: Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC'14). 2014. p. 1547–53.

[14] Salimbajevs A. Creating Lithuanian and Latvian Speech Corpora from Inaccurately Annotated Web Data. In: Calzolari N, Choukri K, Cieri C, Declerck T, Goggi S, Hasida K, et al., editors. Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018 [Internet]. European Language Resources Association (ELRA); 2018.

[15] Povey D, Cheng G, Wang Y, Li K, Xu H, Yarmohamadi M, et al. Semi-orthogonal low-rank matrix factorization for deep neural networks. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2018.

[16] Povey D, Peddinti V, Galvez D, Ghahremani P, Manohar V, Na X, et al. Purely sequence-trained neural networks for ASR based on lattice-free MMI. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2016. p. 2751–5.

[17] Hadian H, Sameti H, Povey D, Khudanpur S. End-to-end speech recognition using lattice-free MMI. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2018.

[18] Paul, D. B., & Baker, J. M. (1992). The design for the Wall Street Journal-based CSR corpus. In Proceedings of the workshop on Speech and Natural Language (pp. 357-362).

[19] Smit P, Virpioja S, Kurimo M. Improved Subword Modeling for WFST-Based Speech Recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2017. p. 2551–5. Available from: http://dx.doi.org/10.21437/Interspeech.2017-103