

USING STACKED TRANSFORMATIONS FOR RECOGNIZING FOREIGN ACCENTED SPEECH

Peter Smit, Mikko Kurimo

Adaptive Informatics Research Centre
Aalto University
firstname.lastname@tkk.fi

ABSTRACT

A common problem in speech recognition for foreign accented speech is that there is not enough training data for an accent-specific or a speaker-specific recognizer. Speaker adaptation can be used to improve the accuracy of a speaker-independent recognizer, but a lot of adaptation data is needed for speakers with a strong foreign accent. In this paper we propose a rather simple and successful technique of stacked transformations where the baseline models trained for native speakers are first adapted by using accent-specific data and then by another transformation using speaker-specific data. Because the accent-specific data can be collected offline, the first transformation can be more detailed and comprehensive, and the second one less detailed and fast. Experimental results are provided for speaker adaptation in English spoken by Finnish speakers. The evaluation results confirm that the stacked transformations are very helpful for fast speaker adaptation.

Index Terms— automatic speech recognition, foreign-accent recognition, cmlr transformation, stacked transformations

1. INTRODUCTION

One of the challenges in automatic speech recognition (ASR) that has great impact in practical applications, is to improve recognition accuracy of foreign accented speech. In the recent years several approaches have been proposed to account for the pronunciation variation, for example, to adapt the pronunciation dictionary [1] or the acoustic model in various ways [2]. Multiple investigations have also been made for cross-lingual adaptation of models, typically by using recordings in the mother tongue of the foreign speaker [3, 4, 5].

In this paper the focus is on recognizing English, when pronounced by native Finnish speakers, an accent sometimes called ‘Finglish’. A foreign accented data set which is collected from Finnish university students is used for adaptation

and another data set, gathered from students and visitors at the University of Edinburgh, for evaluation.

The new simple but efficient approach that we suggest for accented ASR is to split the speaker adaptation into two successive transformations. The first one adapts a general model to accented speech and the second one further to a specific speaker. Thus these *stacked transformations* take advantage of both the diversity and larger amount of non-accented speech and the smaller amount of available accented speech. This differs from the cross-lingual approach [3, 4, 5], because we use accented data instead of data in another language. Also, other acoustic model approaches [2] do not utilize CMLLR transformations estimated with data from multiple speakers.

This work is part of the EMIME project, which aims at personalized speech-to-speech translation (S2ST). To produce synthesized output speech that sounds like the same speaker in different language, it would be convenient to replace part of the speaker-specific data by accent-specific data. Because HMM-based models are utilized in recognition and synthesis, the adaptation framework is the same in both. Thus, in addition of improving foreign accented ASR, this work is also motivated by the goal of improving foreign accented speech synthesis for S2ST [6].

2. STACKED TRANSFORMATIONS

The proposed method of stacked transformations utilizes constrained maximum likelihood linear regression (CMLLR) adaptations [7] to start from a Speaker-Independent (SI) models to first make an intermediate Accent-Dependent (AD) model and then a final Speaker-Dependent (SD) model. In conventional speaker adaptation, an SD model is emulated by adapting a SI model directly to one speaker.

The first step is to adapt the SI model to an AD model. Because there is, relatively, a lot of accented data available, this first transformation can be very detailed, expressing itself in a large number of regression classes. The second step is to adapt the AD model to an SD model. However, because the AD model is already much closer to the SD model than

This work was supported by the Academy of Finland in the project *Adaptive Informatics* and the IST Programme of the European Community, under the FP7 project EMIME (213845).

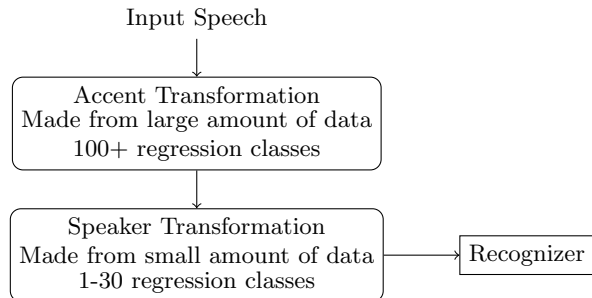


Fig. 1. Diagram of the recognition procedure with stacked transformations. The accent transformation is trained with data from multiple accented speakers. The speaker transformation is trained with only data of the target speaker.

a general SI model, we expect that less SD adaptation data is needed to achieve similar performance than in a conventional speaker adaptation for accented speech.

In recognition, first the accent transformation is applied and after that the speaker transformation, as also is shown in Figure 1.

The expected advantages of stacked transformations are numerous. Firstly, the AD model can be calculated off line, which has no cost in recognition time. Secondly, less speaker-specific data is needed for adaptation. In the case of unsupervised adaptation, the adaptation data can be used more efficiently, because the first ASR pass utilizing the AD model provides a better transcription than the SI model, thus leading to a more accurate second transformation.

If the amount of high quality accented training data is very large, an AD model could also be directly trained, but as there are usually much more and better native training data, an AD transformation is likely to produce better and more detailed models. By pooling the accented data directly with the larger and better quality native data, more robust models could be obtained for accented ASR, but this may not be the best way to take advantage of the two very different data sources.

3. DATA SETS AND MODEL SETUP

3.1. Data sets

For accent adaptation, a data set called *DSP* containing accented English speech was collected from Finnish university students. In total 74 different speakers were recorded with 20 utterances per speaker, totaling 1474 utterances (one speaker only pronounced 14 sentences). 92% of the speakers were male.

The recordings were done using headset microphones in a classroom, thus some soft background chatter and other environment noises are observable in the recordings.

The recorded sentences for each speaker were randomly chosen from two sets. One set contained 200 simple English

sentences from the Herald Tribune database, and the other one 25 Europarl sentences and 100 sentence from the WSJ0 Enrolment and language model test set, which were both more complicated sentences. All sentences were selected at the University of Edinburgh based on phonetic coverage.

For evaluation, we used a data set called here *UED EngF* recorded at the University of Edinburgh [8]. It contains of 14 Finnish speakers (7 Male, 7 Female) speaking each 125 English sentences. Each speakers speaks the same 125 sentences, which are the same Europarl and WSJ0 sentences as used for the *DSP* data. This means that the same sentences can occur in training/adaptation and evaluation¹. However, we do not expect this to affect the results significantly, because the speakers in the two data sets are different. *UED EngF* was recorded in a studio with high-quality equipment and therefore contains no noise. Also the speakers were less accented, compared to the *DSP* data set.

For non-accented reference, we used a data set called here *UED EngN* that has two North American English speakers recorded in exactly the same conditions as *UED EngF*. Because the *DSP* data set contains mainly male speakers, only the Male speakers from the UED data sets were used for evaluation.

For training and evaluating the baseline recognizer, the Wall Street Journal-based corpus (*WSJ*) [9] was used. For training we chose the *WSJ-284* selection (283 speakers, 66 hours of speech) and for evaluation the *WSJ0 20k* evaluation set.

3.2. Models and Recognizer

All models are trained in similar fashion and all recognition experiments are done with the same basic parameters using HTK [10].

The models are all cross-word triphone HMM models with 16 component GMM emission distributions. Silence was modeled in a separate state and a short pause model was used for word breaks. As features, MFCC coefficients with first and second derivative are used. The CMU dictionary and phoneme set were used as lexicon.

The model training was initialized with the flatstart procedure. Triphone tying was used to ensure sufficient training data for each state. In the training procedure the number of Gaussian components was gradually increased until there were 16 components per state.

In recognition the generation of lattices was done with HDecode and a 2-gram language model made from the *WSJ-20k* language model data. As is commonly done with HTK, the lattices were rescored with a 3-gram language model and decoded with the SRILM lattice-tool. There were no out-of-vocabulary words in the evaluation sets.

¹It was not possible to correct this without severely decreasing the size of the *DSP* data set.

	<i>WSJ</i>	<i>DSP</i>	<i>WSJ+DSP</i>	<i>WSJ at</i>	<i>WSJ+DSP at</i>
N_1	3.4	32.9	3.7	3.6	4.1
N_2	9.0	43.8	8.6	8.0	7.9
F_1	49.6	36.0	41.9	37.7	31.9
F_2	39.2	43.7	35.1	32.7	31.5

Table 1. Baseline performance for three different models (columns) on four different evaluation data (N_1 = WSJ0 eval, N_2 = UED EngN, F_1 = DSP, F_2 = UED EngF). Another two models are prepared by adding an accented transformation (“at”) using the DSP data set. The numbers are word error rates (WER) in %.

The stacked transformations are implemented with the so-called ‘parent’ or ‘cascaded’ transformation feature of the HTK-toolkit, which allows to use a sequence of transformations.

4. EXPERIMENTS

4.1. Baseline and accent transformation

The first experiment gives a baseline performance for the collected data and shows the performance of an accented transformation. Three different models are evaluated. One model contains only non-accented (*WSJ*) and one only accented training data (*DSP*). The last model contains both the non-accented and accented data (*WSJ+DSP*).

Four different evaluations are provided for each model. The evaluation sets are: WSJ0 eval (N_1), UED EngF (F_2), and UED EngN (N_2) as described in the previous section. The *DSP* (F_1) data set is too small to be split in a training and evaluation set, so for that data a 10-fold cross validation result is provided. For all other experiments the evaluation and training data are different sets.

The results of this experiment are shown in Table 1 and Figure 2. *WSJ at* and *WSJ+DSP at* are the baseline results for the accented transformation on the model. It is noteworthy that this does not include stacked transformations yet, as no speaker-specific adaptation data and transformation is included. Thus, this is just a different way to utilize the accent-specific part of the training data.

The first thing to note in the results is that the *WSJ* model works very well for the WSJ0 evaluation data (N_1) and the UED EngN data (N_2), but only moderately for the accented data sets. This is all in the line of expectation, because the conditions for the UED EngN data (N_2) only vary slightly from *WSJ*, but the accented data sets have more severe mismatch because of the foreign accent.

Except for the WSJ0 evaluation set (N_1), it seems to be always beneficial to add the Finnish accented data in the training. The reason may be that it adds robustness against the mismatch of recording conditions.

The results for the accent transformation (“at” columns)

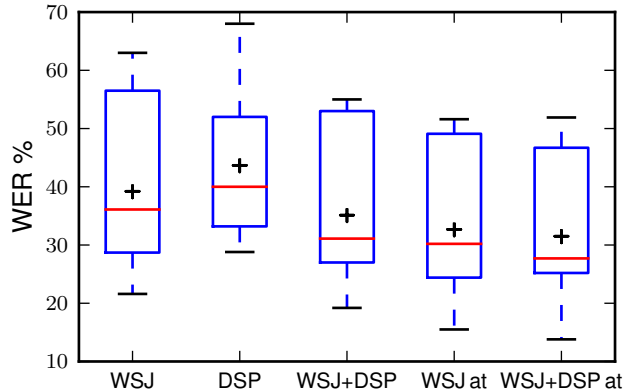


Fig. 2. Distribution of baseline error rate by speaker for all different models on *UED EngF* evaluation. The horizontal line in the middle of each box is the median error, the cross marks the average.

are quite interesting. As expected the transformation significantly improves the recognition compared to the baseline model. Even if the Finnish data was already used in training (*WSJ+DSP*), recycling it for the accent transformation is still useful. It is slightly surprising to see that also UED EngN (N_2) is recognized better when the Finnish accent transformation is used. This could be the effect of having slightly different recording conditions, but it should be investigated further. The WSJ0 evaluation data (N_1) shows the expected result of degraded performance when the accent-specific transformation is used.

4.2. Stacked transformations

As the AD model should already fit to speech with the same accent much better than the other models, we expect that it then needs less speaker-specific adaptation data to obtain further improvements in recognition.

On both the *WSJ* and the *WSJ+DSP* models two experiments are performed. One with only speaker-specific adaptation and one with first an accent transformation and then speaker-specific adaptation (stacked transformations, *st*). In both experiments the amount of speaker-specific adaptation utterances is gradually increased and the development of the average error rate of speakers in UED EngF is shown in Figure 3. The amount of accent-specific data for the accent transformation was kept constant.

The graph in Figure 3 confirms that the stacked transformations are better than the conventional speaker adaptation for both *WSJ* and *WSJ+DSP* models. If the stacked transformations for *WSJ* (*WSJ.st*) is compared to the speaker adaptation of the *WSJ+DSP* model, an identical performance is obtained after 10 adaptation utterances, but for smaller amount of speaker-specific data, the stacked transformations are better. However, applying stacked transformations to the

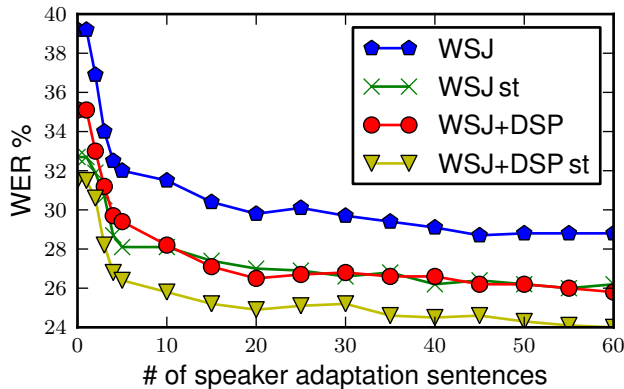


Fig. 3. Development of average WER of speakers in UED EngF for different number of speaker-specific adaptation utterances

WSJ+DSP model (*WSJ+DSP st*) provides clearly the best performance.

The observation that for less than 10 adaptation samples, in particular, the stacked transformations give a superior performance compared to the conventional speaker adaptation, indicates that this technique would be useful in situations where very little adaptation data is available.

5. CONCLUSION AND FUTURE WORK

This paper shows that for foreign accented speech it is beneficial to use stacked transformations for speech recognition. Compared to employing speaker adaptation to a model that is not accent-specific, the relative improvement is from 16% (with no speaker-specific adaptation data) to 9% (with 60 adaptation utterances). If the base model already includes some training data of the target accent, the improvements range from 10% (with no speaker-specific adaptation data) to 7% (with 60 adaptation utterances).

In future work we will check the benefits of the stacked transformations using data that is not only accent-specific. This means that from the accent-specific data, the speakers included in the first transformation could be further selected based on gender, age, or location information. Automatic methods for selecting similar speakers will also be studied. In both cases the stacked transformation might also be useful for non-accented speech recognition.

The UED data sets have been collected with the purpose of detecting cross-lingual speaker similarity. The listener perspective on the degree of accent of each speaker could be compared with the ASR results shown here, to see if there exists any correlation between the degree of accent and the effectiveness of the accent-specific transformation.

As mentioned in the introduction, this method might be also beneficial for speech synthesis. Because there is a correlation between the amount of adaptation data available and

the quality of resulting transformed voice [11], the stacked transformations presented in this paper could also be applied there for obtaining a better initial transformation.

6. REFERENCES

- [1] Tao Chen Chao Huang and Eric Chang, “Accent issues in large vocabulary continuous speech recognition,” *International Journal of Speech Technology*, vol. 7, no. 2-3, pp. 141–153, 2004.
- [2] Z. Wang, T. Schultz, and A. Waibel, “Comparison of acoustic model adaptation techniques on non-native speech,” in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP’03). 2003 IEEE International Conference on*. IEEE, 2003, vol. 1.
- [3] Reima Karhila and Mikko Kurimo, “Unsupervised cross-lingual speaker adaptation for accented speech recognition,” in *to appear in SLT*, 2010.
- [4] Stefanie Aalburg and Harald Hoega, “Foreign-accented speaker-independent speech recognition,” in *Inter-speech*, 2004, pp. 1465–1468.
- [5] Katarina Bartkova and Denis Jouviet, “On using units trained on foreign data for improved multiple accent speech recognition,” *Speech Comm.*, vol. 49, no. 10-11, pp. 836 – 846, 2007.
- [6] M. Wester, J. Dines, M. Gibson, H. Liang, Y. Wu, L. Saheer, S. King, K. Oura, P. N. Garner, W. Byrne, Y. Guan, T. Hirsimäki, R. Karhila, M. Kurimo, M. Shannon, S. Shiota, J. Tian, K. Tokuda, and J. Yamagishi, “Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project,” in *7th ISCA Speech Synthesis Workshop*, 2010.
- [7] M.J.F. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.
- [8] M. Wester, “The emime bilingual database,” Tech. Rep. EDI-INF-RR-1388, University of Edinburgh, September 2010.
- [9] D.B. Paul and J.M. Baker, “The design for the Wall Street Journal-based CSR corpus,” in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.
- [10] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book*, Citeseer, 1997.
- [11] J. Yamagishi and T. Kobayashi, “Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training,” *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 2, pp. 533, 2007.