UNSUPERVISED CROSS-LINGUAL SPEAKER ADAPTATION FOR ACCENTED SPEECH RECOGNITION

Reima Karhila and Mikko Kurimo

Adaptive Informatics Research Centre
Aalto University School of Science and Technology, Finland
firstname.lastname@tkk.fi

ABSTRACT

In this paper we present investigations on how the acoustic models in automatic speech recognition can be adapted across languages in unsupervised fashion to improve recognition of speech with a foreign accent. Recognition systems were trained on large Finnish and English corpora, and tested both on monolingual and bilingual material. Adaptation with bilingual and monolingual recognisers was compared. We found out that recognition of foreign accented English with help of Finnish adaptation training data from the same speaker was not improved significantly. However, the recognition of native Finnish using foreign accented English adaptation data was improved significantly.

Index Terms— automatic speech recognition, multilingual acoustic modelling, cross-lingual speaker adaptation

1. INTRODUCTION

The recent penetration of speech technology to a wide variety of languages has increased the importance of multilingual and adaptive automatic speech recognition (ASR) and synthesis (text-to-speech, TTS). One of the key problems there is how to effectively adapt the ASR and TTS models for new voices when there is little or no training data for a particular speaker in a particular language. Also, when offering ASR-based services to a linguistically varying population, another problem in the case of ASR is how to recognise speech with a foreign accent.

The ASR of accented speech is a difficult problem in many ways. Even if speech is fluent, there may be a severe mismatch with the available training data, which is usually from native speakers. Previously, ASR of foreign accented speech has been tried with acoustic models (AM) adapted to accents [1, 2] or with Pronunciation Dictionary Adaptation [3]. Both methods require large amounts of data or knowledge about the foreign language in advance. These methods can be combined and further improved with more traditional speaker-adaptation methods. It has been shown that components of Gaussian mixture models can be used to model pronunciation variations [4], but only within the variants present in the training data. Our goal is to improve ASR in a foreign language with adaptation methods so that we don't need any previous data from foreign speakers of that language.

In speaker adaptation the recogniser learns the personal characteristics of the voice of the speaker using adaptation data. When there is little or no training data for the speaker in the target language, one can try to learn the voice characteristics from her or his speech in another language. In this paper we have studied and evaluated Cross-Lingual Speaker Adaptation (CLSA) for large-vocabulary continuous speech recognition (LVCSR). In CLSA, adaptation training data for the target speaker is not available in the target language, but only in another language. Recently, the progress in CLSA of synthetic voices has been promising for the HMM-based TTS using rather similar adaptation techniques that were originally developed for ASR [5]. Thus, it is interesting to study how the same techniques help CLSA of ASR models, though collecting a small amount of adaptation data in the target language might still provide higher gains in the ASR performance.

One of the key problems in CLSA is segmenting the source language training data into phonemes that match well enough to the HMMs for the target language so that the data can be used to adapt the ASR or TTS models. When there is little adaptation data or its segmentation is inaccurate, it is important to enable robust adaptation.

In this paper we first describe our approaches to CLSA in Section 2. We compared these CLSA methods in LVCSR performance for Finnish speakers when they speak English. The baseline acoustic models were English models trained on the WSJ0 corpus. As the reference performance we have used ASR with intra-lingual adaptation, where the models are adapted on the English speech of the target speakers. The models that apply adaptation in cross-lingual setting were adapted using only the Finnish speech of the target speakers. The experimental results are reported in Section 3.

We then reversed the experiment conditions and ran the

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).

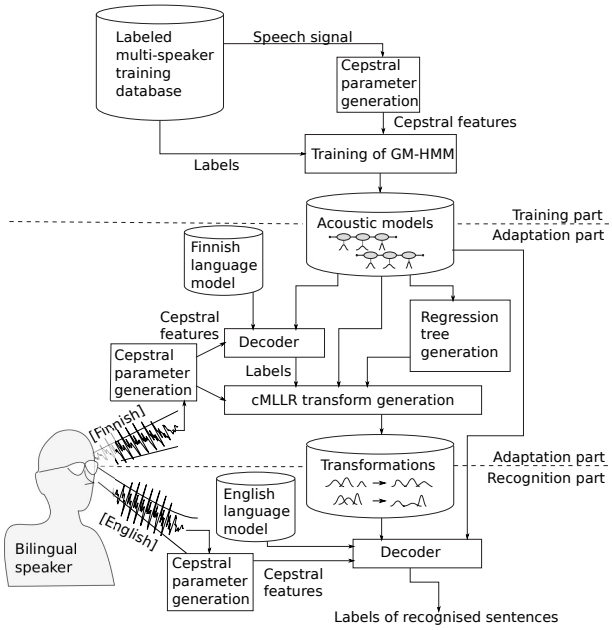

Fig. 1. *Cross-lingual adaptation in multilingual ASR.*

comparison of CLSA methods using the target speakers' non-native English speech to adapt their native Finnish models. These results are reported in Section 4, followed by discussion and conclusions in Sections 5 and 6.

2. CROSS-LINGUAL SPEAKER ADAPTATION (CLSA)

In our approach, we try to use native speech from a foreign speaker directly to enhance the ASR of foreign accented speech. In CLSA, speaker-specific linear transforms are trained from one language and applied to the AMs of another language. In this study, two approaches are compared.

In the *Transformation Sharing* (TS) approach (Fig. 1) the transformations are trained on the same AM set as is used for recognition. The transformations are computed for the models that appear in the adaptation speech data, and shared with models in the target language. This requires a multilingual AM set. In transformation sharing, our main concern is to use correct transformations for each triphone model. For this, the Gaussian components of the triphone models are clustered using the Euclidean distance. 32 clusters are created, and when enough adaptation training data is not available, a regression tree is used to pool the clusters as much as needed.

In the *State Mapping* (SM) [5] approach we calculate a mapping between the AMs of the two languages. The labels of the target speakers' adaptation training data are then mapped to the test recognition language, and the recogniser is adapted using these labels. State mapping can be done with any pair of monolingual recognisers. We can simplify the system by using probabilistic state mapping. This mapping is acquired by recognising the adaptation training data with a phoneme loop recogniser [6]. The triphone phoneme loop uses the target language AMs to recognise the source language adaptation data. Thus we get an approximate phonetic labelling of the data as if it were in another language. For this approach, only a single monolingual recogniser is required. The same regression tree method is used here as in the transformation sharing approach.

3. EXPERIMENT 1: IMPROVING RECOGNITION OF FOREIGN ACCENTED ENGLISH

The main goal of this study was to find out whether CLSA can improve ASR performance in recognising foreign accented English. A secondary objective was to investigate if and how sharing training data across corpora affects CLSA. For these purposes, four LVCSR systems, ML-Sep, ML-Mix$_0$, ML-Mix$_{13}$ and ML-Mix$_{100}$ with different data sharing approaches were trained and tested with mono- and bilingual speech data.

3.1. Training, adaptation, and evaluation data

Bilingual recognisers ML-Mix$_0$, ML-Mix$_{13}$ and ML-Mix$_{100}$ are trained with the American English WSJ0 corpus and the Finnish Speecon corpus. The close microphone recordings were used from both corpora. ML-Sep has two separate acoustic models, an English AM is trained with WSJ0 and a Finnish one that is not used in this experiment. The English training set Tr_{en} is the SI-84 data set. The Speecon corpus had been divided into a training set Tr_{fi} and evaluation test set Bas_{fi}. The English baseline test set Bas_{en} is the si-et-05 test set of WSJ0.

The bilingual test data was originally recorded for testing the adaptation of TTS voices from three male native Finnish speakers. The Finnish test set Exp_{fi} consists of 125 sentences (100 sentences from Speecon corpus prompts and 25 sentences from the European parliament corpus texts). The English test set Exp_{en} consists of 86 sentences (26 WSJ0 enrolment sentences and 60 WSJ0 language model test set sentences). These sets are selected so that the out-of-vocabulary rate is 0. Additionally, subsets of the WSJ0 enrolment set and of Exp_{en}, San_{en} and San_{fi}, were used as a "sanity check" to verify that the results of the adaptation experiments were within reasonable boundaries. The data sets are listed in Table 1.

3.2. Training multilingual recognisers

All the speech corpora were preprocessed identically. The computed features were 12 Mel-Cepstral coefficients and power. Energy normalisation and channel-specific cepstral mean subtraction (CMS) normalisation were used. First and second derivatives of the features were appended to create 39

Table 1. *Databases used for training and testing recognisers.*

Set	Lang.	Accent	Speakers / Sentences	Notes
Tr_{fi}	Finnish	Native	330 / 16963	Speecon training set
Tr_{en}	English	Native	83 / 7123	WSJ0 SI-84
Bas_{fi}	Finnish	Native	40 / 1118	Speecon baseline test set
Bas_{en}	English	Native	8 / 333	WSJ0 si_et_05
Exp_{fi}	Finnish	Native	3 / 375	Experiment test set
Exp_{en}	English	Foreign	3 / 264	Experiment test set
San_{for}	English	Foreign	3 / 78	Subset of Exp_{en}, identical prompts with San_{nat}
San_{nat}	English	Native	8 / 208	Subset of WSJ0 Enrolment

dimensional feature vectors. Here the preprocessed corpora are assumed to be acoustically close enough, so that any clustering of phones between them would depend more on the phonetic properties than on the recording conditions.

All the recognisers use cross-word triphone, 3-state left-to-right HMM models with Gaussian mixtures of 16 components as emitting states. The 39-phoneme CMU set was used for English and Speecon's 23 phoneme set for Finnish. Also one long and one short model were included for silence.

The single Gaussian, diagonal covariance monophone models were initialised by flat-starting and then copied to triphone models. Triphone models were tied (clustered) using a decision tree with phonetic questions to ensure an adequate amount of training data per model. The full Gaussians mixture for each model was trained by gradually increasing the number of components.

Semi-tied covariance (STC) transforms [7] are added to the model set. In all systems except one there is one STC transform for the centre phone of each triphone. The exception is mix_{100}, which has a global STC transform because of its state clustering system. Both STC geometries give significant improvement over plain diagonal covariance systems in one-pass recognition. Additionally, STC transformed model sets are further trained by speaker-adapted training (SAT).

It was found out that the chosen STC transformation grouping does not go well with the cMLLR adaptation transforms, and thus also the results using plain "vanilla" models without STC are included.

3.3. Cross-lingual model combination

Multilingual ASR systems capable of CLSA can be constructed by combining model sets from several languages. The multilingual model set can be condensed by clustering models across languages. To obtain a clustering in feature space, the Gaussian representations of the emitting HMM states can be used as a basis. The similarity metric in this work is based on the Kullback-Leibler (KL) divergence, because it is well suited for calculating distances between Gaussian distributions. The distance between two phone models was calculated by adding together the KL divergences of the emitting Gaussian distributions of each respective HMM state of the two models. The distances in both directions are added together to get a proper distance metric.

Three different approaches were investigated in training multilingual recognisers:

A **Sep**: The baseline recogniser consisted of separate AMs for English and Finnish, and no cross-lingual clustering. In this case, only the English model set is used.

B $\textbf{Mix}_0$ and $\textbf{Mix}_{13}$: Recognisers were trained using both corpora. The triphones were tied with help of a decision tree for phonetic questions shared across languages. This tree was constructed by combining phonetically identical context-questions from both languages. Mix_0 uses no cross-lingual clustering. In Mix_{13}, 26 models of the 13 closest cross-lingual monophone pairs were merged pairwise. A new model set was then trained by pooling acoustic data from both languages for these models right from the start. In context-clustering, each phone is considered to be a member of all the languages in which it appears.

The distance of the monophones of the two languages was calculated from the monophone models of the baseline recognisers according to the described distance metric. A representation of the phoneme distances is shown in Figure 2. This figure shows that Finnish and English phones are mixed well with each other with phonetically similar phones grouped together.

C $\textbf{Mix}_{100}$: The third approach was to cluster triphones of the baseline recognisers across languages after triphone tying. Several thresholds were tested, and the results are shown for the best system with 100 triphones from both languages merged. Triphone states are combined across languages instead of complete triphones. Otherwise, it would be necessary to either create a new transition for the combined model, which would need more training data, or to make both models use a transition from one language only.

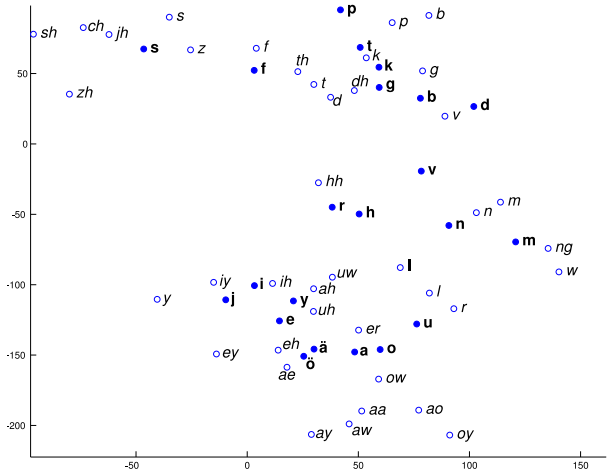

Fig. 2. *Multi-factor dimensionality reduction graph of phones of WSJ and Speecon corpora. Finnish phones (Speecon) are marked with solid circles and bold text, and English phones (CMU) as circle outlines and italics.*

3.4. Phoneme loop label generation

Simplifying the version of adaptation scheme in [6], a recognition network is built with context-dependent models of one language, and used to recognise speech data from a second language. Unconstrained by language models the resulting labels look like gibberish, but give the best fit alignment of the models of one language on the data of another language. These aligned labels can then be used to train adaptation transforms.

3.5. Baseline performance

Baseline test results are shown as an indicator of the quality of the ASR systems. The English ASR performance in baseline test Bas_{en} is shown in Table 2. As the generation of adaptation labels for CLSA requires also a Finnish recogniser, the results for a test with Bas_{fi} are shown in the same table.

Single pass and two-pass results are reported. In two-pass recognition, first pass result labels are used to generate speaker-specific adaptation transforms. The adaptation method used in this work is constrained maximum likelihood linear regression (cMLLR) done with 3 block-diagonal 13x13 transform matrices. All of the available data in one language is used to create the speaker-specific transforms. The English LM is WSJ's 5k vocabulary models and the Finnish LM is a general 44k vocabulary morphed model. In all tests, a word lattice is first generated using a 2-gram LM. These lattices are then rescored with a larger n-gram, where $n = 3$ for English and $n = 10$ for Finnish [8].

3.6. Adaptation results

The following tests were made:

Table 2. *Baseline performance for English by word error rate and for Finnish by letter error rate.*

Set	Adapt	Model type	Error rates for ASR System Sep	mix_0	mix_{13}	mix_{100}
Bas_{en}	no	diag	5.6	7.7	7.6	6.3
Bas_{en}	cMLLR	diag	3.7	4.4	4.7	3.3
Bas_{en}	no	STC	4.8	6.9	7.2	6.2
Bas_{en}	cMLLR	STC-SAT	4.3	4.4	4.7	3.3
Bas_{fi}	No	Diag	4.2	4.1	4.1	4.1
Bas_{fi}	cMLLR	Diag	3.4	3.5	3.5	3.5
Bas_{fi}	No	STC	3.7	3.7	3.7	3.7
Bas_{fi}	cMLLR	STC-SAT	3.5	3.7	3.6	3.2

1. Single-pass recognition of English test set Exp_{en},
2. Two-pass recognition; 1st pass on English test set Exp_{en}, 2nd pass on the same with cMLLR adaptation generated with 1st pass labels,
3. Transformation sharing adaptation; 1st pass on Finnish test set Exp_{fi}, followed by a recognition pass on English test set Exp_{en} with a cMLLR adaptation trained from Exp_{fi}
4. State mapping adaptation; 1st pass on Finnish test set Exp_{fi} with a phoneme recogniser, followed by a recognition pass on English test set Exp_{en} with cMLLR adaptation trained from Exp_{fi}

The results are shown in Table 3. The best results that were statistically significantly better (matched pair test) than best single pass results (36.8 WER) have been shaded.

Table 3. *Error rates in Experiment 1 CLSA tests for* exp_{en} *using word error rate.*

Test	Adapt data	Model type	Error rates for ASR System Sep	mix_0	mix_{13}	mix_{100}
1.1	-	diag	41.6	46.6	48.6	42.3
1.2	En	diag	29.6	35.9	34.2	31.0
1.3	Fi TS	diag	-	60.2	54.7	38.0
1.4	Fi SM	diag	35.7	49.3	49.6	35.7
1.1	-	STC	36.8	42.4	41.3	38.3
1.2	En	STC-SAT	31.5	37.1	36.4	29.3
1.3	Fi TS	STC-SAT	-	102.5	88.6	38.9
1.4	Fi SM	STC-SAT	37.1	74.4	66.4	35.5

With CLSA techniques, the only statistically significant improvement was between the 1st pass and adapted 2nd pass results of the ML-Mix_{100} recogniser. Compared to the best single pass recognition of ML-Sep STC model set, any improvement was however statistically insignificant, and often results only got worse. 2-pass recognition on English data gives a 15-23% reduction in word error rate compared to the

Table 4. *Word error rates of enrolment sentences, the same 26 sentences from 8 speakers of WSJ0 set and 3 speakers of bilingual set, using the WSJ20k 3-gram LM. No adaptation is used.*

Test set	Model Type	Error rates for ASR System Sep	mix_0	mix_{13}	mix_{100}
San_{nat}	STC	11.4	15.6	14.8	11.3
San_{for}	STC	41.1	46.3	45.6	42.6

unadapted systems, whereas the SM CLSA gives a 4-7% reduction compared to the unadapted systems.

Because the ASR results for foreign English (Table 3) are so much worse than for the native English (Table 2), some additional comparisons were made. Table 4 shows the performance for the san_{for} enrolment sentences in the recorded bilingual data and the san_{nat} enrolment sentences of WSJ0. These results are in line with tests in accented speech recognition with no adaptation techniques [9, 10] where the increase in error rate has been some hundreds of percents for differently accented speech.

As is apparent from all of the test results, merging of models between English and Finnish with ML-Mix_0 and ML-Mix_{13} approaches has been highly detrimental for English ASR performance. The ML-Mix_{100} system does not do as well as the ML-Sep in unadapted performance, but gives better results when adaptation is used.

4. EXPERIMENT 2: IMPROVING RECOGNITION OF NATIVE FINNISH

The goal of this study was to find out whether CLSA investigated in Experiment 1 works symmetrically. A secondary objective was again to examine effects of data sharing.

4.1. Differences to Experiment 1

Here the languages are switched, so we investigate improving native Finnish recognition with the help of foreign English from the same speakers. The multilingual recognisers were the same as in Experiment 1. The ML-Sep system is trained from the Speecon corpus. The test sets are used but in different order. No sanity check has been done, as the results of the tests are in line with current Finnish ASR results [8]. Otherwise, the arrangements are identical to Experiment 1.

4.2. Adaptation results

The following tests were made:

1. Single-pass recognition of Finnish test set Exp_{fi},
2. Two-pass recognition; 1st pass on Finnish test set Exp_{fi}, 2nd pass on the same with cMLLR adaptation generated with 1st pass labels,
3. Transformation sharing adaptation; 1st pass on English test set Exp_{en}, followed by a recognition pass on Finnish test set Exp_{fi} with a cMLLR adaptation trained from Exp_{en}
4. State mapping adaptation; 1st pass on English test set Exp_{en} with a phoneme recogniser, followed by a recognition pass on Finnish test set Exp_{fi} with cMLLR adaptation trained from Exp_{en}

The results are shown in Table 5. The results that were statistically significantly better (matched pair test) than best single pass results (3.4 LER) have been shaded.

Table 5. *Error rates in Experiment 2 CLSA tests for* $\exp_{fi}$ *using letter error rate.*

Test	Adapt type	Model type	Error rates for ASR System Sep	mix_0	mix_{13}	mix_{100}
2.1	-	diag	4.2	5.3	4.6	4.2
2.2	Fi	diag	2.4	2.3	2.3	2.3
2.3	En TS	diag	-	3.0	2.6	2.6
2.4	En SM	diag	2.8	3.1	2.6	2.8
2.1	-	STC	3.4	4.0	3.5	3.4
2.2	Fi	STC-SAT	2.7	2.7	2.8	2.2
2.3	En TS	STC-SAT	-	4.2	3.6	2.6
2.4	En SM	STC-SAT	3.4	3.3	3.4	2.8

When recognising Finnish using adaptation transforms computed from foreign English, all systems show significant improvements. 2-pass recognition gives at best 29-43% reduction in letter error rate compared to the unadapted systems. The worse the baseline performance, the bigger the gain from adaptation. CLSA methods give 18-29% improvement in recognition rate compared to the unadapted systems, with no statistically significant difference between the state mapping and transformation sharing approaches.

As is apparent from the baseline and adaptation tests in Section 3.5, the merging of models for multilingual ASR has little degrading effect for ASR of native Finnish.

5. DISCUSSION

Several phenomena behave asymmetrically between the two languages. First, merging of models between languages degrades English ASR performance far more than Finnish ASR performance. Secondly, adaptation of the native language works better, even when the error is quite low to start with. Thirdly, no degrading behaviour was perceived in the Finnish CLSA tests.

The most obvious thing is the asymmetrical behaviour of CLSA. Why this yields significant improvement in native language but not in foreign language is something that requires some thought. The behaviour of MLLR transform generation

when the error rate is high is a question that can be bypassed, as we do see improvement in intra-lingual case.

One reason for this would be acoustic differences between training and test data. The recording conditions between the training and test speech are different, but still the native Finnish sentences recorded with the same speakers using the same equipment did not pose any problems in Finnish ASR. Also, MLLR transforms adapt also to microphone and background noise. This leaves the way the recorded speakers speak English as the main factor in the increase in error.

Another reason might be the unsuitability of the English model set for these particular speakers. The English acoustic models were trained only on 84 speakers, whereas the Finnish ones were trained with 310 speakers, and the test data was spoken by native Finns. A more probable issue is the phone model usage mismatch between native and accented English. Experiment 2 shows that speaker characteristics can be learned over languages, but Experiment 1 suggests that the foreign pronunciation mismatch cannot be learned by adapting from the native language. Either this phenomenon is too complicated to be modelled by a simple linear transform or CLSA adapts the wrong models, whereas intra-lingual adaptation affects the right model groups. This issue could maybe be countered by creating more robust average voice models, that take into account systematic pronunciation errors by foreign speakers.

An interesting thing with the generally meagre performance of the multilingual systems is that there is significant improvement in the adapted ASR performance when using the Mix_{100} model set. However, it should still be tested whether this improvement arises from the different transform geometry of its STC model set.

6. CONCLUSIONS

We have shown that CLSA works in speech recognition, even if only in helping recognise native speech. Although this reverse direction of speaker adaptation may not have many applications in practise, our observations may help to understand better the current problems in multilingual ASR.

We also found out that the use of probabilistic state mapping across monolingual recognisers is a valid way of generating phoneme labels for adaptation and is as good or better as using multilingual recognisers to generate the labels. This, and the mediocre performance of the mixed systems, means that clustering acoustic models across languages does not seem to be worthwhile using the tested methods.

Possible future investigations include repeating the tests with other language pairs, including the same languages the other way around, with a database of native American English and foreign Finnish. Unfortunately acquiring such database is not trivial. Another interesting investigation would be building more robust English models and including foreign English in the training data.

7. REFERENCES

[1] Katarina Bartkova and Denis Jouvet, "On using units trained on foreign data for improved multiple accent speech recognition," *Speech Comm.*, vol. 49, no. 10-11, pp. 836 – 846, 2007.

[2] Stefanie Aalburg and Harald Hoege, "Foreign-accented speaker-independent speech recognition," in *Interspeech*, 2004, pp. 1465–1468.

[3] 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.

[4] Murat Saraçlar, Harriet Nock, and Sanjeev Khudanpur, "Pronunciation modeling by sharing Gaussian densities across phonetic models," *Computer Speech & Language*, vol. 14, no. 2, pp. 137 – 160, 2000.

[5] Y.-J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," in *Interspeech*, 2009, pp. 528–531.

[6] M. Gibson, "Two-pass decision tree construction for unsupervised adaptation of HMM-based synthesis models," in *Interspeech*, 2009, pp. 1791–1794.

[7] M.J.F. Gales, "Semi-tied covariance matrices for hidden markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 272–281, 1999.

[8] T. Hirsimäki, J. Pylkkönen, and M Kurimo, "Importance of high-order n-gram models in morph-based speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 4, pp. 724–732, 2009.

[9] Qin Yan and S. Vaseghi, "A comparative analysis of UK and US english accents in recognition and synthesis," in *ICASSP*, 2002, pp. 413–416.

[10] Ayako Ikeno, Bryan Pellom, Dan Cer, Ashley Thornton, Jason M. Brenier, Dan Jurafsky, Wayne Ward, and William Byrne, "Issues in recognition of Spanish-accented spontaneous English," in *IEEE/ISCA Workshop on Spontaneous Speech Processing and Recognition*, 2003.