

# UNSUPERVISED CROSS-LINGUAL SPEAKER ADAPTATION FOR HMM-BASED SPEECH SYNTHESIS USING TWO-PASS DECISION TREE CONSTRUCTION

Matthew Gibson<sup>1</sup>, Teemu Hirsimäki<sup>2</sup>, Reima Karhila<sup>2</sup>, Mikko Kurimo<sup>2</sup>, William Byrne<sup>1</sup>

<sup>1</sup>Cambridge University Engineering Department, Trumpington Street, Cambridge CB2 1PZ, U.K.

<sup>2</sup>Helsinki University of Technology, Adaptive Informatics Research Centre, P.O.Box 5400, Finland

## ABSTRACT

This paper demonstrates how *unsupervised* cross-lingual adaptation of HMM-based speech synthesis models may be performed without explicit knowledge of the adaptation data language. A two-pass decision tree construction technique is deployed for this purpose. Using parallel translated datasets, cross-lingual and intralingual adaptation are compared in a controlled manner. Listener evaluations reveal that the proposed method delivers performance approaching that of unsupervised *intralingual* adaptation.

**Index Terms**— HMM-based speech synthesis, unsupervised speaker adaptation, cross-lingual.

## 1. INTRODUCTION

Cross-lingual (or interlingual) speaker adaptation is defined as the adaptation of acoustic models associated with one language, the *target language*, using adaptation data uttered in a different language, the *source language*. Recent work [1, 2] has addressed the task of supervised cross-lingual adaptation for HMM-based speech synthesis (or text-to-speech, TTS). This work used TTS models of both source and target languages, and defined a phoneme or state-level mapping between the source and target language acoustic models. This mapping was deployed during cross-lingual adaptation to translate the source transcription to a target language phoneme or state sequence.

Techniques similar to those described above rely upon the availability of both source and target language TTS models, and the mapping mechanism between these models must be established prior to adaptation. The work described in this paper quantifies the value of such prior knowledge by adopting an alternative approach which requires no knowledge of the source language acoustic model (or source language) and its relationship to the target language acoustic model. Further, this alternative technique is applied to the task of *unsupervised* cross-lingual adaptation.

Using parallel translated adaptation datasets recorded by the same speaker, *intralingual* (within-language) and cross-lingual adaptation are compared in a controlled manner. Listener evaluations reveal that the proposed unsupervised cross-

lingual adaptation delivers performance approaching that of unsupervised intralingual adaptation.

The paper is structured as follows. Sections 2 and 3 respectively discuss the idea and implementation of the cross-lingual adaptation technique. Section 4 describes the experimental procedure used to evaluate the technique and Section 5 analyses the results of the evaluation. Section 6 summarises the contributions of this work.

## 2. UNSUPERVISED CROSS-LINGUAL ADAPTATION OF SPEECH SYNTHESIS MODELS

The cross-lingual adaptation technique used in this work treats the source language adaptation data as if it were uttered in the target language. Target language acoustic models and a phoneme-loop grammar are used to recognise the adaptation data, thus mapping it onto a phoneme sequence  $\hat{p}$  in the target language. Subsequently, the estimated phoneme sequence  $\hat{p}$  is used as the reference sequence by the speaker adaptation algorithm. This process is similar to standard approaches to unsupervised intralingual adaptation. Note that to avoid language-specific constraints, no dictionary or language model is used during recognition.

Several advantages are associated with this technique. Firstly, a pre-trained source language acoustic model is not required. Secondly, no previously learned mapping between source and target language acoustic models is necessary. Thirdly, the technique may be applied even when the source language is unknown. However, it may be argued that automatic mapping of source language speech to a target language phoneme sequence is suboptimal. The impact of this automatic mapping is measured in Section 4.

To perform unsupervised adaptation in this way, it is necessary to perform automatic speech recognition (ASR) with the target TTS models to obtain transcription  $\hat{p}$ . This is problematic because the acoustic models typically used in HMM-based speech synthesis are not easily integrated into the ASR search procedure. This, in turn, is because the context-dependent acoustic models used in HMM-based speech synthesis represent suprasegmental information (e.g. syllabic stress). However, these models, henceforth referred to as *full context* models, may be constructed such that they

may be mapped to triphone acoustic models suitable for ASR. This technique, introduced in [3], is summarised in Section 3.

### 3. TWO-PASS DECISION TREE CONSTRUCTION

Full context models are clustered using a decision tree to enable robust estimation of their parameters. By imposing constraints upon the decision tree structure, multiple-component triphone mixture models may be derived from single-component full context models. This constrained decision tree construction process is illustrated in Figure 1.

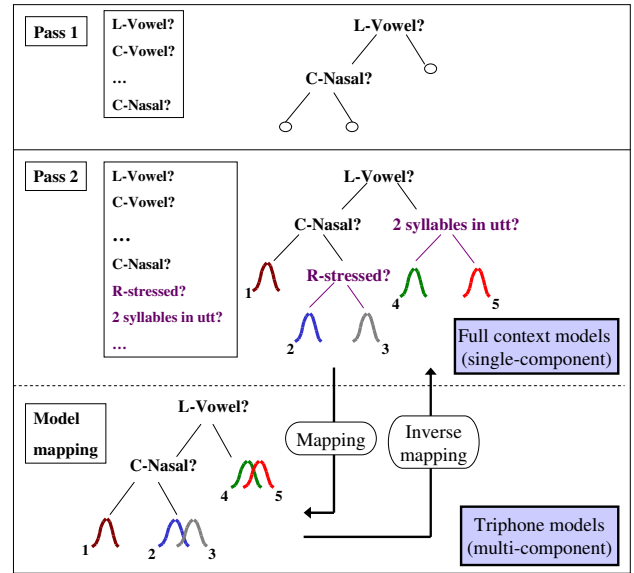
The first stage, indicated as Pass 1 in Figure 1, uses only questions relating to left, right and central phonemes to construct a phonetic decision tree. This decision tree is used to generate a set of tied triphone contexts, which are easily integrated into the ASR search. Pass 2 extends the decision tree constructed in Pass 1 by introducing additional questions relating to suprasegmental information. The output of Pass 2 is an extended decision tree which defines a set of tied full contexts. After this two-pass decision tree construction, single component Gaussian state output distributions are estimated to model the tied full contexts associated with each leaf node of the extended decision tree. These models are then used for speech synthesis.

A mapping from the single-component full context models to multiple-component triphone models is defined as follows. Each leaf node of the extended decision tree has a unique ‘triphone ancestor’ node, namely its ancestor leaf node of the Pass 1 decision tree. Each set of Gaussian components associated with the same ‘triphone ancestor’ is grouped as components of a multiple component mixture distribution to model the context defined by the ‘triphone ancestor’. The derived triphone models are illustrated at the bottom of Figure 1. The weight of each mixture component is calculated from the occupancies associated with components of the Pass 2 leaf node contexts. The inverse mapping from triphone models to full context models is obtained by associating each Gaussian component with its original full context. Given this mapping between full context and triphone models, unsupervised adaptation of full context acoustic models may be achieved via adaptation of triphone models, as described below.

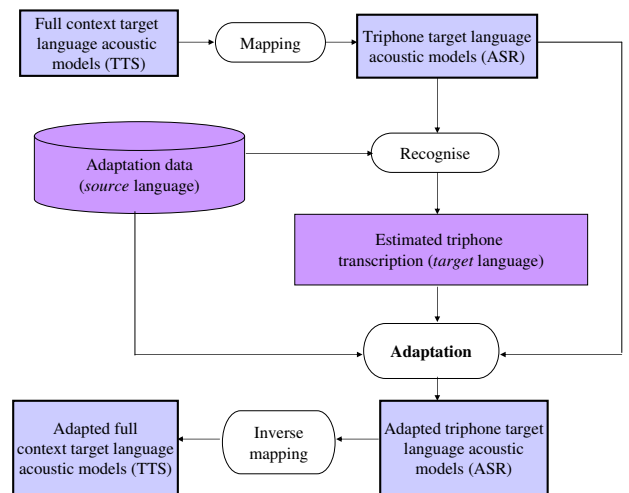
As illustrated in Figure 2, triphone models derived from full context models are used to estimate a triphone-level transcription of source language adaptation data. This estimated transcription is then used to adapt the triphone models. The adapted triphone models are subsequently mapped back to full context models using the inverse mapping.

### 4. EXPERIMENTS

Full context English average voice models are estimated using speaker adaptive training (SAT, [4]) and the Wall Street Journal (WSJ) SI84 dataset. Acoustic features used



**Fig. 1.** Two-pass decision tree construction. Mapping functions permit sharing of full context models for TTS and triphone models for ASR.



**Fig. 2.** Unsupervised cross-lingual adaptation of full context target language acoustic models.

are STRAIGHT-analysed Mel-cepstral coefficients [5], fundamental frequency, band aperiodicity measurements, and the first and second order temporal derivatives of all features. The acoustic models use explicit duration models [6] and multi-space probability distributions [7]. Decision trees (one per state and stream combination) are constructed using the two-pass technique of Section 3. Adapted TTS systems are

derived from the average voice models using the adaptation method described in Section 3 and constrained maximum likelihood linear regression. Speech utterances are generated from models via feature sequence generation [8] and resynthesis of a waveform from the feature sequence [5].

#### 4.1. Adaptation and evaluation datasets

The adaptation datasets comprise 94 utterances from a corpus of parallel text of European parliament proceedings [9]. English and Finnish versions of this dataset are recorded in identical acoustic environments by a native Finnish speaker also competent in English. Statistics relating to these datasets are provided in Table 1. The evaluation dataset comprises En-

Language	# utterances	# minutes	# words
English	94	12.3	1546
Finnish	94	10.9	1066

**Table 1.** Europarl adaptation datasets.

glish utterances (distinct from the adaptation utterances) from the same Europarl corpus.

#### 4.2. Evaluation

The following systems are evaluated.

- System A: average voice.
- System B: unsupervised cross-lingual adapted.
- System C: unsupervised intralingual adapted.
- System D: supervised intralingual adapted.
- System E: vocoded natural speech.

System B is the result of applying unsupervised cross-lingual adaptation to the average voice models using the Finnish adaptation dataset. System C results from unsupervised adaptation using the English adaptation dataset. System D is identical to System C with the exception that the correct transcription is used during adaptation. System E analyses and resynthesises the evaluation utterances using STRAIGHT[5].

All systems were evaluated by listening to synthesised utterances via a web browser interface, as used in the Blizzard Challenge 2007. The evaluation comprised four sections. In the first pair of sections, listeners judged the naturalness of an initial set of synthesised utterances. In the second pair of sections, listeners judged the similarity of a second set of synthesised utterances to the target speaker’s speech. Four of the target speaker’s natural English utterances were available for comparison. Each synthetic utterance was judged using a five point psychometric response scale, where ‘5’ and ‘1’ are respectively the most and least favourable responses.

Twenty-four native English and sixteen native Finnish speakers conducted the evaluation. Different Latin squares were used for each section to define the order in which systems were judged. Each listener was assigned a row of each

Latin square, and judged five different utterances per section, each synthesised by a different system.

## 5. RESULTS

Figure 3 summarises listener judgements of ‘similarity to target speaker’ and ‘naturalness’ using boxplots [10] while Table 2 displays the average mean opinion scores (MOS) of these judgements for each system in the columns labelled ‘av’. Analysis of these judgements by listener native language is provided in the columns labelled ‘En’ and ‘Fi’, respectively denoting English and Finnish.

Sys	Source lang.	Sup?	MOS similarity			MOS naturalness		
			En	Fi	av	En	Fi	av
A	-	-	1.2	1.1	1.1	2.3	2.4	2.3
B	Fi	N	2.3	1.5	2.0	2.4	2.4	2.4
C	En	N	2.6	1.7	2.2	2.6	2.7	2.7
D	En	Y	2.7	2.0	2.4	2.5	2.8	2.6
E	-	-	4.6	4.6	4.6	3.7	4.1	3.8

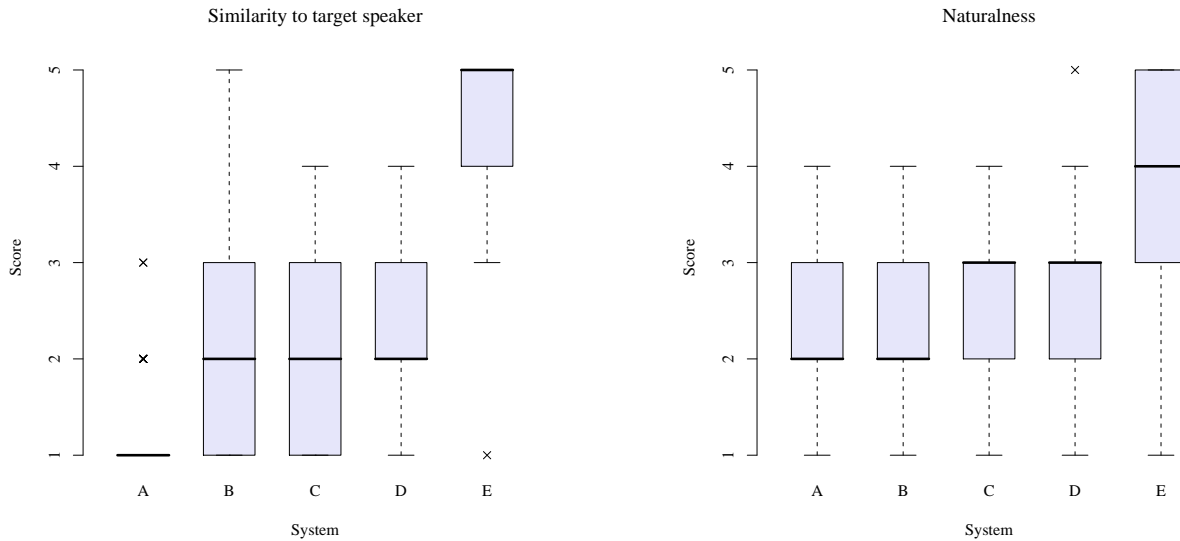
**Table 2.** Mean opinion scores of evaluated systems.

A difference between two systems is deemed significant if the Bonferoni-corrected pairwise Wilcoxon signed rank test [10] discovers significance at the 95% confidence level.

#### 5.1. Discussion

The average similarity to the target speaker given by adapted systems (B, C and D) are all significantly greater than that observed for the unadapted System A. A significant difference is also observed between the average similarity of adapted systems B (2.0) and D (2.4). The similarity score for system D is a reasonable upper limit on the performance of *supervised* cross-lingual adaptation since system D uses the correct transcription of the acoustic data and deploys no potentially sub-optimal mapping between source and target language acoustic models.

The similarity score for system C is a reasonable upper limit on the performance of *unsupervised* cross-lingual adaptation for the given imperfect adaptation data transcription. System B yields an average similarity (2.0) approaching that of System C (2.2). This difference may be due to the additional knowledge used by System C i.e. prior knowledge of source language acoustics and a perfect mapping from source to target language acoustic models. In this case the small performance degradation measures the value of such knowledge. A second possible explanation should be kept in mind, however. The target speaker’s characteristics may change when speaking his non-native English. When adapting using native Finnish speech, such alterations are not observed, and so possibly not captured.



**Fig. 3.** Listener opinion scores for similarity to target speaker and naturalness.

Small increases in naturalness are noted between the adapted systems (B, C and D) and the unadapted system A. This demonstrates that the improvements in similarity discussed above do not compromise the naturalness of the synthetic speech.

Lastly, note that, with respect to speaker similarity, Finnish and English judges display similar patterns. However, on average, Finnish listeners ascribe lower scores. Further analysis is required to explain this observation. Influential factors may include familiarity with the target speaker and cultural differences.

## 6. CONCLUSION

This paper has measured the value of source language knowledge upon the task of unsupervised cross-lingual adaptation of HMM-based synthesis models. A cross-lingual adaptation technique which uses no such knowledge delivers performance approaching that of unsupervised intralingual adaptation. Future work should measure how well this technique generalises across different speakers and language pairs. Future work may also determine if significant differences exist between unsupervised intralingual and cross-lingual adaptation, and explain any such differences.

## 7. ACKNOWLEDGEMENTS

We are very grateful to the organizers of the Blizzard Challenge for providing experimental evaluation scripts. This research was partially funded by the European Communitys Seventh Framework Programme (FP7/2007-2013), grant agreement 213845 (EMIME).

## 8. REFERENCES

- [1] Y. Wu, S. King, and K. Tokuda, "Cross-lingual speaker adaptation for HMM-based speech synthesis," in *Proceedings ISCSLP*, 2008.
- [2] Y. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," in *Proceedings Interspeech*, 2009.
- [3] M. Gibson, "Two-pass decision tree construction for unsupervised adaptation of HMM-based synthesis models," in *Proceedings Interspeech*, 2009.
- [4] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Audio, Speech & Language Processing*, vol. 17(1), pp. 66–83, 2009.
- [5] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.
- [6] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," in *Proceedings Interspeech*, 2004.
- [7] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. & Syst.*, vol. E85-D(3), pp. 455–464, 2002.
- [8] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proceedings ICASSP*, 2000.
- [9] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in *Machine Translation Summit*, 2005.
- [10] R. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," in *Proceedings of the Blizzard challenge workshop*, 2007.