interspeech 2008 incorporating SST 08

Cross-lingual Portability of MLP-Based Tandem Features – A Case Study for English and Hungarian

László Tóth[1], *Joe Frankel*[2], *Gábor Gosztolya*[1], *Simon King*[2]

[1]Research Group on Artificial Intelligence, Hungarian Academy of Sciences, Szeged, Hungary
[2]Centre of Speech Technology Research, University of Edinburgh, UK

tothl@inf.u-szeged.hu, joe@cstr.ed.ac.uk, ggabor@inf.u-szeged.hu, Simon.King@ed.ac.uk

Abstract

One promising approach for building ASR systems for less-resourced languages is cross-lingual adaptation. Tandem ASR is particularly well suited to such adaptation, as it includes two cascaded modelling steps: feature extraction using multi-layer perceptrons (MLPs), followed by modelling using a standard HMM. The language-specific tuning can be performed by adjusting the HMM only, leaving the MLP untouched.

Here we examine the portability of feature extractor MLPs between an Indo-European (English) and a Finno-Ugric (Hungarian) language. We present experiments which use both conventional phone-posterior and articulatory feature (AF) detector MLPs, both trained on a much larger quantity of (English) data than the monolingual (Hungarian) system. We find that the cross-lingual configurations achieve similar performance to the monolingual system, and that, interestingly, the AF detectors lead to slightly worse performance, despite the expectation that they should be more language-independent than phone-based MLPs. However, the cross-lingual system outperforms all other configurations when the English phone MLP is adapted on the Hungarian data.

Index Terms: automatic speech recognition, multilayer perceptrons, articulatory features, cross-lingual

1. Introduction

Over the past couple of decades, multi layer perceptrons (MLPs) have become part of mainstream automatic speech recognition (ASR) in the tandem [1] approach. In tandem ASR, the MLPs provide non-linear mappings of the feature space for use in standard hidden Markov model (HMM) systems. The addition of the MLP-derived tandem features have been shown to yield significant benefits on a variety of domains and languages (see, e.g. [2, 3, 4]).

Typically, the MLPs are trained to estimate the phone posteriors, though other target values may also be of use – the only requirement being that the MLP-based transformation should improve modelling in the subsequent Gaussian mixture model (GMM) data description step. An alternative choice of MLP target is that of articulatory features (AFs). A number of authors have proposed AF recognition as a reasonable first step of speech recognition [5, 6], the main arguments being that they should provide a better account of pronunciation variability, and also they are more language universal than phones. The set of AF classifier multilayer perceptron (MLP) neural nets we use here was trained on 2000 hours of telephone speech for the 2006 Johns Hopkins Summer Workshop [7, 8, 9].

The aim of this paper is to study the cross-lingual portability of tandem features based on phone and articulatory feature classifier MLPs. One of the core motivations for such research is the difference in the amount of training data available for the various languages. In our case, the articulatory feature MLPs for the source language, English, were trained on 2000 hours of speech, while we have only 7 hours of data available in the target language, Hungarian. We intend to explore the possibilities for transferring the well-trained feature MLPs between languages, despite the clear phonetic differences. Furthermore, we expect that the tandem architecture will be particularly suitable for cross-lingual adaptation, as it consists of two cascaded statistical modelling steps.

In Sections 2-3 we develop a tandem recognizer using only the Hungarian training data. In Section 4, we investigate cross-lingual porting by simply taking the English MLPs without any modifications and retraining only the HMM component on the language-specific data. We then extend this to the situation where the tandem MLPs are also adapted. We present our conclusions in Section 5.

2. Experimental setup

2.1. Data

As a Hungarian training and test corpus we used the MTBA Hungarian Telephone Speech Database [10]. To our knowledge, this is the only commercially available Hungarian telephone speech corpus of a reasonably large size. The primary data block of the corpus contains sentences that were read out loud by 500 speakers. The sentences are relatively long (40-50 phones per sentence) and were selected in order to give good coverage of all the most frequent Hungarian phone pairs. Each speaker reads 16 sentences, with each sentence occurring three times in the database. An additional set of 4 words were recorded by each speaker in order to cover the phone combinations that were missing from the sentences. Recordings were made via both mobile and landline phones, and the speakers were selected to give an age and gender distribution matching that of the Hungarian population. All the sentences and words were manually time-aligned at the phone level using a set of 58 phonetic symbols. Infrequently occurring allophones were collapsed with other units giving a final working set of 52 phones.

This manually processed part of the database altogether comprises of seven hours of speech data in 8000 wave files. To form training and testing data sets, all files where the transcript indicated the presence of noise (both background and speaker-originated) were first removed. The remaining 6935 files were divided into two blocks in a ratio of approximately 4 to 1, resulting in training and testing sets of 408 and 91 speakers, respectively. No language model for a large vocabulary word transcription task was available, so our experiments consisted of

Accepted after peer review of full paper

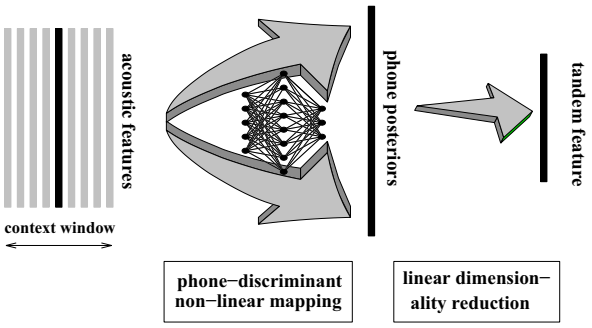

Figure 1: Schematic of phone-based tandem ASR.

phone recognition on the test utterances.

To complement these results, we performed isolated word recognition experiments on a further block of the database consisting of 500 distinct city names, each one spoken by a different caller. A pronunciation dictionary was constructed for them using an automatic phonetic transcription routine. From the 500 recordings, only 431 were used in the evaluation, as the rest contained significant noise or were misread by the caller.

2.2. Tandem ASR

All experiments use a standard HMM recognition system based on HTK [11]. Phone models are monophone 3-state left-to-right HMMs with 9 diagonal-covariance Gaussian components per state. As a baseline feature set, 12 perceptual linear prediction (PLP) cepstra plus energy were derived from the acoustic signal from 25ms windows at a frame shift of 10ms. These were mean and variance normalized on a per-speaker basis, and then expanded with the usual Δ and $\Delta\Delta$ coefficients, resulting in a spectral feature vector of 39 components.

The original tandem ASR [1] uses multi-layer perceptrons (MLPs) to provide a non-linear transformation of the acoustic features. Multiple frames of acoustic parameters are fed into an MLPs, which have been trained to perform phone classification. Rather than interpreting the outputs as phone class posteriors as in hybrid artificial neural network (ANN) / HMM ASR [12], they are transformed as described below by logarithm transformation and dimensionality reduction, then treated as *discriminative features*, to be modelled by the Gaussian mixture models in a standard HMM system. This process is outlined in Figure 1.

Here we follow the standard tandem approach of taking a context window of 9 frames (central frame plus 4 frames of left and right context) of PLP cepstra as input to a 3-layer MLP. The hidden units have sigmoid activation functions, and there is a softmax over the output units. Once generated, the tandem features are subjected to dimensionality reduction as described below and appended to the the 39 PLP coefficients to give the input feature vector of the HMM.

MLP training was performed using the `Quicknet` tools developed at ICSI, following standard procedures including using cross validation data to determine the MLP learn rate and convergence. These are freely available for download at `www.icsi.berkeley.edu/Speech/qn.html`.

2.3. Dimensionality reduction of MLP outputs

Tandem MLPs are typically trained to perform phone class discrimination, and given the softmax activation function, the outputs can be interpreted as phone posteriors [12]. However, in order to improve subsequent modelling with Gaussian mixture models (GMMs), a logarithm transform is first applied to expand the dynamic range. In addition, attempts are made to decorrelate the features, as the subsequent Gaussian modelling step assumes zero cross-covariance values. One possibility is to use principal component analysis (PCA), and like Ellis et al, retain all components [1]. Alternatively, dimensionality reduction can be introduced into this step, for example by selecting the largest PCA components which together account for 95% of the data variation [9]. We present experiments which evaluate both of these possibilities. Furthermore, as the discrete cosine transform (DCT) is known to have a decorrelating effect quite close to that of PCA [13], we also evaluate this method as comparison.

3. Monolingual experiments

A Hungarian tandem MLP was optimized on the training data described in Section 2.1. The 9 frame input window with 39 PLP cepstra gave 351 input units, and the number of hidden units was set to 500 according to accuracy on a held-out cross validation set. There were 52 output units, one for each phone class. In addition, an MLP was trained with 4800 hidden units to provide a network with an equal number of free parameters to that of the cross-lingual MLP described in Section 4 below, even though it was expected that there would be an excessive number of free parameters for the amount of training data.

Table 1 shows the word and phone recognition results obtained using the PLP features, the MLP-based features with various types of post-processing transformations, and two feature sets consisting of PLPs concatenated with PCA-postprocessed MLP-based features. The results suggest that adding tandem features to the PLPs yields performance improvements. Choosing the number of PCA dimensions in order to account for 95% of the data variation meant retaining only 12 and 11 components for the 500 and 4800-unit MLPs respectively. Whilst the word accuracies are comparable, the phone recognition results show that PCA-based decorrelation with no dimensionality reduction gives the best results.

In addition, it appears that training the large tandem MLP did not result in over-fitting, and in fact gave slight performance improvements. Therefore all the following evaluations and comparisons will be based on the configuration with 4800 hidden units, and PCA decorrelation with no dimensionality reduction.

4. Cross-lingual experiments

Training a large MLP from scratch requires significant amounts of data, and the availability of time-aligned labels. In this section we present experiments which explore methods for using MLPs trained on English for the task of Hungarian ASR. Such techniques should prove useful in deploying ASR systems for languages for which limited resources are available. As one would reasonably expect that the success of such cross-lingual modelling depends greatly on the similarity of the two languages, we first briefly summarize the main phonetic differences between English and Hungarian.

4.1. Differences between English and Hungarian

Hungarian is a Finno-Ugric language, and one of the few modern European languages that do not belong to the Indo-European language family. This means that there are several

Feature set	500 hidden neurons			4800 hidden neurons		
	WACC	PhCORR	PhACC	WER	PhCORR	PhACC
PLP	96.1%	59.0%	52.4%	96.1%	59.0%	52.4%
MLP(log)	95.1%	60.1%	53.6%	96.3%	62.0%	55.7%
MLP(log+DCT)	94.4%	60.2%	54.6%	96.1%	62.0%	57.1%
MLP(log+PCA$_{95\%}$)	95.6%	53.5%	50.9%	96.3%	53.4%	51.2%
MLP(log+PCA$_{all}$)	95.6%	64.1%	60.4%	96.8%	65.7%	62.1%
PLP + MLP(log+PCA$_{95\%}$)	94.7%	62.7%	57.2%	97.2%	62.7%	57.6%
PLP + MLP(log+PCA$_{all}$)	**96.3%**	**66.5%**	**61.5%**	**97.5%**	**67.6%**	**62.6%**

Table 1: Word and phone recognition performance (WACC: word accuracy, PhACC: phone accuracy, PhCORR: phone percent correct) for the Hungarian tandem with MLPs of 500 and 4800 hidden neurons, using the PLP features, the MLP-based features (with four types of post-processing steps), and two concatenated PLP plus tandem feature sets. The best result in each column is shown in bold face.

significant differences between the phonetics of English and Hungarian [14]. Their consonant sets are relatively similar, the biggest mismatches being the dental fricatives of English and the palatal affricates of Hungarian, which are both missing from the other language. However, there are also several allophonic differences in those consonants that are present in both languages (for example, voiceless stop consonants are never aspirated in Hungarian). Another feature of Hungarian is that practically all the consonants can be geminated.

There are much bigger differences in the vowel systems. Even the similar monophthongs take slightly different positions in the vowel triangle, while some of them are missing from the other language (e.g. /æ/ from Hungarian, /ø/ and /y/ from English). Even more significantly, in Hungarian there are no diphthongs (apart from dialects and sloppy speech) and unstressed vowels do not get reduced (or only to a much lesser extent than in English). Similarly with the consonants, most vowels have a long and a short variant, and their duration acts as a distinctive feature.

4.2. Crosslingual MLPs

In this work we present and compare three different approaches for constructing and training MLPs for cross-lingual tandem feature extraction. We consider AF classification MLPs trained on 2000 hours of English continuous telephone speech (CTS) data, phone classification MLPs trained on 100 hours of English meetings data, and lastly we adapt the English meeting phone MLPs to Hungarian phone classification.

4.2.1. English articulatory feature MLPs

We might expect that with articulatory features (AF) being based on a universal set of physical constraints, AF MLPs will naturally transfer between languages. Here we use the set of AF classification MLPs which were trained as part of the Johns Hopkins 2006 summer workshop [7]. The structure and training of the MLPs follow the same approach as described in Section 2, though with the multi-valued discrete AFs as the targets rather than phones. A separate feature detector MLP was trained for each of the 8 feature groups given in Table 2. For use in the tandem system, the outputs of the 8 MLPs were concatenated to give a 64-dimensional vector.

4.2.2. English phone MLPs

Phone classification MLPs were trained on over 100 hours of audio collected in instrumented meeting rooms at a number institutes, with participants' speech recorded on headset-mounted microphones. These institutes include the International Computer Science Institute (ICSI), National Institute for Standards and Technology (NIST), Carnegie Mellon University Interactive Systems Laboratory (ISL), plus partners of the Augmented Multiparty Interaction (AMI) project. The hidden layer of the MLP consisted of 4800 units, and the output layer 46, corresponding to the size of the phone set.

Feature	**Values**
Place	labial, labio-dental, dental, alveolar, post-alveolar, velar, glottal, rhotic, lateral, none, silence
Degree/manner	vowel, approximant, flap, fricative, closure, silence
Nasality	+, -, silence
Glottal state	voiced, voiceless, aspirated, silence
Rounding	+, -,silence
Vowel	aa, ae, ah, ao, aw1, aw2, ax, ay1, ay2, eh, er, ey1, ey2, ih, iy, ow1, ow2, oy1, oy2, uh, uw, not-a-vowel, silence
Height	very high, high, mid-high, mid, mid-low, low, nil, silence
Frontness	back, mid-back, mid, mid-front, front, silence

Table 2: The set of 8 multi-levelled articulatory feature groups as used in the Johns Hopkins 2006 summer workshop.

4.2.3. Adapted phone MLPs

The input-hidden layer of an MLP can be considered to be performing speech pattern extraction and dimensionality expansion, whilst the hidden-output layer provides a mapping to the particular task the MLP is being used for. Therefore, a Hungarian phone MLP was trained using the following initialization scheme: the initial weights and biases of the connections between the 351 unit input to 4800 unit hidden layer were created from the English phone MLP's weights and biases, while randomly-initialized weights were used for the hidden to 52 unit output layer connections. The training on the Hungarian data was then performed exactly the same way as with the monolingual system in Section 3.

4.3. Results and discussion

The cross-lingual experiments followed the same procedure as the mono-lingual experiments described above, only varying the tandem MLPs. Preliminary experiments showed that as with the mono-lingual systems, dimensionality reduction via PCA retains too few components and hence reduces performance.

Feature set: PLP	WACC	PhCORR	PhACC
+ Hungarian phone MLP	97.5%	67.6%	62.6%
+ English AF MLP	97.2%	65.9%	61.8%
+ English phone MLP	96.5%	66.3%	62.2%
+ Adapted phone MLP	**98.1%**	**69.3%**	**64.9%**

Table 3: Word and phone recognition performance obtained with the mono-lingual and cross-lingual MLP-based tandem features. The best result in each column is shown in bold face.

Therefore the full decorrelated outputs were used.

Table 3 shows the ASR results from combining tandem features from the 3 cross-lingual MLPs with PLPs as above. For reference we include the Hungarian result from Table 1. These results show that the English phone and AF MLPs lead to similar performance to the Hungarian phone MLPs, and that AF-based tandem features do not transfer more readily than phones between languages. It is worth noting that in addition to the language mismatch, the potential benefit of these MLPs, which have been trained on 1 or 2 orders of magnitude more data than the Hungarian phone MLPs, may also be at a disadvantage due to domain and channel differences.

However, the adapted phone MLPs lead to improvements in word and phone accuracy over all other systems, a result which is shown to be significant in a paired t-test with $p < 0.01$.

Similar cross-language investigations by Çetin at al. [8] also found the same AF MLPs to be slightly less portable than the phone MLPs. They report experiments in which the cross-lingual systems were not able to match the performance of the monolingual one. In this work the MLPs were ported to Mandarin, which is further from English than Hungarian.

Stolcke et al. present cross-lingual experiments with Mandarin and Arabic [15]. By combining the English-trained MLPs with their baseline systems they obtain significant improvements for both languages. Unfortunately, their monolingual systems use no MLP features, so their results are not conclusive regarding the monolingual versus cross-lingual tandem issue.

5. Conclusions

In this paper we investigated the porting of tandem features between languages, to examine whether the increased amount of training data available for the cross-lingual setup will yield improvements despite the mismatch between languages. We found that using MLPs trained on English data gave similar performance to the MLP trained on Hungarian, the target language, in spite of their access to a much greater amount of training data.

However, cross-lingual adaptation of the MLP from English to Hungarian gave a significant improvement over the performance of the MLP trained only on the available Hungarian data. This indicates that for tandem adaptation, the feature extractor MLPs should be adapted in addition to the HMM Gaussian components. Future work will include extending the MLP adaptation techniques, and also looking at methods of domain normalization in order to improve the performance without adapting large MLPs to the target language.

6. References

[1] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*, 2000, pp. 1635–1638.

[2] Q. Zhu, A. Stolcke, B. Chen, and N. Morgan, "Incorporating Tandem/HATs MLP features into SRI's conversational speech recognition system," in *Proc. EARS RT-04F Workshop*, 2004.

[3] M. Karafiát, F. Grézl, P. Schwarz, L. Burget, and J. Černocký, "Robust heteroscedastic linear discriminant analysis and LCRC posterior features in meeting data recognition," in *Proc. MLMI*, 2006, pp. 275–284.

[4] M.-Y. Hwang, W. Wang, X. Lei, J. Zheng, O. Çetin, and G. Peng, "Advances in mandarin broadcast speech recognition," in *Proc. Interspeech*, 2007, pp. 2613–2616.

[5] K. Kirchhoff, G. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, pp. 303–319, 2000.

[6] S. Stueker, F. Metze, T. Schultz, and A. Waibel, "Integrating multilingual articulatory features into speech recognition," in *Proc. Eurospeech*, 2003, pp. 1033–1036.

[7] J. Frankel, M. Magimai-Doss, S. King, K. Livescu, and O. Çetin, "Articulatory feature classifiers trained on 2000 hours of telephone speech," in *Proc. Interspeech*, 2007, pp. 2485–2488.

[8] O. Çetin, M. Magimai-Doss, K. Livescu, A. Kantor, S. King, C. Bartels, and J. Frankel, "Monolingual and crosslingual comparison of tandem features derived from articulatory and phone MLPs," in *Proc. ASRU*, 2007, pp. 36–41.

[9] O. Çetin, A. Kantor, S. King, C. Bartels, M. Magimai-Doss, J. Frankel, and K. Livescu, "An articulatory feature-based tandem approach and factored observation modeling," in *Proc. ICASSP*, 2007.

[10] K. Vicsi, L. Tóth, A. Kocsor, and J. Csirik, "MTBA – a hungarian telephone speech database (in hungarian)," *Híradástechnika*, vol. LVII, no. 8, pp. 35–43, 2002. [Online]. Available: http://alpha.tmit.bme.hu/speech/hdbMTBA.php

[11] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.2)*, Cambridge University Engineering Department., 2002.

[12] H. Bourlard and N. Morgan, *Connectionist Speech Recognition – A Hybrid Approach*. The Kluwer Int. Series in Engineering and Computer Science, 1994.

[13] A. Akansu and R. Haddad, *Multiresolution Signal Decomposition*. Academic Press, 1992.

[14] T. Szende, *Handbook of the International Phonetic Association*. Cambridge University Press, 1999, ch. Illustrations of the IPA: Hungarian, pp. 104–107.

[15] A. Stolcke, F. Grezl, M.-Y. Hwang, X. Lei, N. Morgan, and D. Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *Proc. ICASSP*, 2006, pp. 321–324.