CROSS-DOMAIN AND CROSS-LANGUAGE PORTABILITY OF ACOUSTIC FEATURES ESTIMATED BY MULTILAYER PERCEPTRONS

Andreas Stolcke[1,2] *František Grézl*[2,5] *Mei-Yuh Hwang*[3]
Xin Lei[3] *Nelson Morgan* [2,4] *Dimitra Vergyri*[1]

[1]SRI International, Menlo Park, CA, USA
[2]International Computer Science Institute, Berkeley, CA, USA
[3]University of Washington, Seattle, WA, USA
[4]University of California, Berkeley, CA, USA
[5]Brno University of Technology, Brno, Czech Republic

ABSTRACT

Recent results with phone-posterior acoustic features estimated by multilayer perceptrons (MLPs) have shown that such features can effectively improve the accuracy of state-of-the-art large vocabulary speech recognition systems. MLP features are trained discriminatively to perform phone classification and are therefore, like acoustic models, tuned to a particular language and application domain. In this paper we investigate how portable such features are across domains and languages. We show that even without retraining, English-trained MLP features can provide a significant boost to recognition accuracy in new domains within the same language, as well as in entirely different languages such as Mandarin and Arabic. We also show the effectiveness of feature-level adaptation in porting MLP features to new domains.

1. INTRODUCTION

Traditionally, the feature extraction front ends of speech recognition systems have been designed and optimized by hand and heuristics, sometimes guided by knowledge of the human auditory system [1]. A more satisfying approach would be to perform feature extraction in a data-driven manner, according to an objective function closely related to the recognition task, or at least to train the front end's parameters according to such a criterion. This was achieved in the *Tandem* approach to hybrid connectionist/HMM modeling [2], based on prior work in neural-network-based acoustic modeling [3]. The Tandem approach consists of training a multilayer perceptron (MLP) to perform phone posterior estimation at the frame level, based on traditional (e.g., perceptual linear prediction—PLP) features, and then to use these posteriors (possibly after further transformations, such as log and dimensionality reduction) as features in a standard Gaussian mixture-based hidden Markov model (HMM) recognizer. The immediate advantage of this approach is trainability of the features, alleviating the modeling burden on Gaussian models by what amounts to a nonlinear transformation optimized for phone discrimination. Further advantages are that the MLP can be given multiple frames of input features, thereby enabling modeling of a larger temporal window. In a further development, the Tandem approach has been extended to include multiple MLPs, based on different input features and operating at different time scales, whose output posteriors are combined into a single, more accurate, and more robust estimate [4, 5]. We have recently shown that such multiple MLP-based feature extractors can give significant error reductions even in complex, state-of-the-art large vocabulary recognition systems [6].

Since the feature extraction algorithm is now trained from data, the acoustic model of the recognizer is effectively factored into two components: one or more MLPs for feature estimation, and the traditional Gaussian mixture models. This raises the question of how portable (domain-independent) the feature extraction is, since generalization is a perennial problem with standard acoustic models. In the extreme case, we can ask if such features trained on one language are suitable for another. If the features did not generalize to different data sources, trainability would seem more of a liability than an advantage for many applications. Conversely, if it turned out that features, once trained, give an advantage on unseen kinds of speech we would have a powerful tool for leveraging training data across domains and languages.

In this paper we address the portability question by using a set of MLP features previously trained on English conversational telephone speech, and found very effective when tested on matched data [6]. We tested these features on a very different English recognition task, multiparty meetings, which also allowed us to investigate simple approaches to feature (in addition to model) adaptation. Finally, we tested generalization to Mandarin and Arabic conversational telephone speech.

2. RECOGNITION SYSTEM

We briefly outline SRI's English conversational telephone speech (CTS) recognition system. This was the system for which MLP features were originally trained and optimized.

2.1. MLP features

In our system, MLP features augment (rather than replace) standard Mel-frequency cepstral coefficient (MFCC) and PLP features. MLPs are trained by taking various snapshots of the time-frequency plane as input. The MLP posteriors can later be combined for higher accuracy. We have found that posteriors from MLPs focusing on information derived from long time chunks of 500 ms can be effectively combined with posteriors from MLPs focusing on shorter-duration chunks of 100 ms. The combined posterior goes through further transformation including log, principal

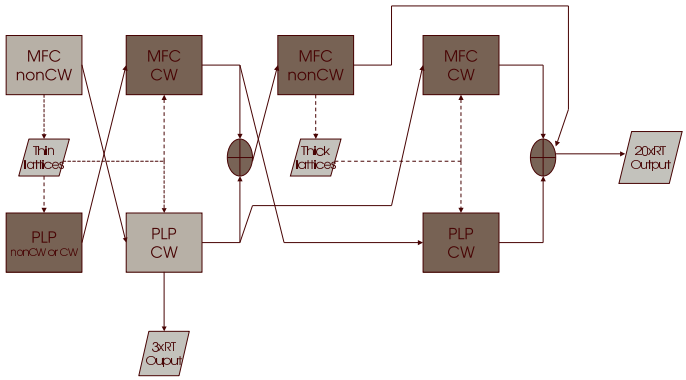

Fig. 1. SRI CTS recognition system. Rectangles represent decoding steps. Parallelograms represent decoding output (lattices or 1-best hypotheses). Solid arrows denote passing of hypotheses for adaptation or output. Dashed lines denote generation or use of word lattices for decoding. Crossed ovals denote confusion network system combination. The two decoding steps in light gray can be run by themselves to obtain a "fast" system.

component analysis (PCA), and truncation in the way described in [7], and is then concatenated with the traditional features such as MFCC or PLP to form the augmented feature vector for Gaussian modeling.

For both types of MLPs, the output targets are the 46 phones used in the SRI CTS recognition system. The MLP focusing on medium-term information takes 9 consecutive frames of PLP features, as well as their first and second deltas, as inputs. To extract long-term information, we use a variant of the Temporal Patterns (TRAPs) MLP architecture [4] called Hidden Activation TRAPs (HATs) [8]. A HATs feature extractor consists of two stages of MLPs. The first stage extracts phonetically discriminant information from 500 ms of critical band energies, while the second stage merges this information and produces phone posteriors. The phone posteriors from both systems are merged on a per-frame basis using a weighted average, where the weights are the inverse entropy of the phone posteriors coming from the corresponding system [9].

The feature MLPs were trained separately for male and female speakers, on a total of 1800 hours of CTS data from the Fisher and Switchboard corpora. Each of the 4 MLPs (male/female, Tandem/HATs) had about 8 million parameters. Various optimizations and heuristics were employed [6] to achieve an acceptable training time (5 weeks).

2.2. Decoding architecture

The recognition architecture is depicted in Figure 1. An "upper" (in the figure) tier of decoding steps is based on MFCC features; a parallel "lower" tier of decoding steps uses PLP features [1]. The outputs from these two tiers are combined twice using word confusion networks (denoted by crossed ovals in the figure). Except for the initial decodings, the acoustic models are "cross-adapted" to the output of a previous step from the respective other tier using maximum-likelihood linear regression (MLLR) [10]. The initial decoding steps in each tier also use MLLR, though with a a phone-loop model as reference.

Lattices are generated initially to speed up subsequent decoding steps. The lattices are regenerated once later to improve their accuracy, after adapting to the outputs of the first combination step. The lattice generation steps use noncrossword (nonCW) triphone models, while decoding from lattices uses crossword (CW) models. The final output is the result of a three-way system combination of MFCC-nonCW, MFCC-CW, and PLP-CW models. Each box in the diagram corresponds to a complex recognition step involving a decoding run to generate either lattices or N-best lists, followed by a rescoring of these outputs with higher-order language models, duration models, and a pause language model [11].

Table 1. Results (WER) on RT-04F English CTS development and evaluation sets. "Fast" refers to the 3xRT 2-stage recognition system, whereas "Full" denotes results with the full evaluation system.

	RT-04F devtest		RT-04F eval	
Features	Fast	Full	Fast	Full
Baseline	18.2	17.2	21.7	20.3
w/MLP features	16.8	15.5	20.0	18.3

The acoustic models employed in decoding use standard normalization techniques: cepstral mean and variance normalization, vocal tract length normalization (VTLN) [12], heteroscedastic linear discriminant analysis (LDA) [13, 14], and speaker-adaptive training based on constrained MLLR [15]. All acoustic models are trained discriminatively using the minimum phone error (MPE) criterion [16]. Acoustic models are trained on about 2300 hours of Fisher and Switchboard CTS data. The baseline language models (LMs) are bigrams (for lattice generation), trigrams (for lattice decoding), and 4-grams (for lattice and N-best rescoring). The CTS in-domain training materials are augmented with data harvested from the Web, using a search engine to select data that is matched for both style and content [17].

The entire system runs in about 13 times real time on a 3.4 GHz Intel Xeon processor. For many scenarios it is useful to use a "fast" subset of the full system consisting of just two decoding steps (the light-shaded boxes in Figure 1); this fast system runs in about 3 times real time (3xRT) and exercises all the key elements of the full system except for confusion network combination.

2.3. In-domain results

As described in [6], it is suboptimal to augment all acoustic models in a complex recognition like the one in Figure 1 with MLP features. In our system, we obtained best results by using MLP features only in the top tier of recognition steps, that is, with MFCC-based models. PLP-based (lower tier) models are left unchanged.

Using this setup, we evaluated the effect of adding MLP features on the NIST RT-04F English CTS development and evaluation testsets, each consisting of 72 telephone conversations. Results are summarized in Table 1. The relative word error rate (WER) reduction using the full system is identical on both testsets, 9.9% (2.0% absolute reduction on the evaluation set).

3. CROSS-DOMAIN EXPERIMENTS

We now investigate how the MLP features perform in another English recognition task. The task was the recognition of multiparty meetings, as required for the NIST RT-05S evaluation. We focus here on the recognition from individual head-mounted microphones (the "IHM" condition in the NIST evaluation). The meeting recognition system was an adapted version of the CTS system described above [18]. The meeting speech recordings were down-

Table 2. Results (WER) on IHM meeting recognition using various sets of acoustic models and features. Columns 2 and 3 indicate whether the Gaussian models or the MLP features were adapted to the meeting domain. "None" in column 3 indicates that MLP features were not used at all, whereas "no" means that CTS-trained MLPs were used.

	Gaussians adapted?	MLP adapted?	RT-04S	RT-05S
a	no	no	28.9	28.6
b	no	yes	28.4	27.0
c	yes	none	29.4	28.6
d	yes	no	28.6	26.9
e	yes	yes	28.3	26.2

sampled to 8 kHz to match the bandwidth of the CTS models. All CTS acoustic models were adapted to the available IHM meeting training data (about 100 hours), using the maximum a posteriori (MAP) criterion, with a combination weight empirically optimized on held-out data. The language models were retrained using available meeting transcripts, as well as background data from CTS and Web data. An audio preprocessing step was added to eliminate crosstalk from background speakers. The recognition architecture was otherwise the same as shown in Figure 1.

In addition to adapting the acoustic (Gaussian) models, we experimented with *feature adaptation*. This was accomplished by performing three incremental MLP training iterations on the phone alignments of the meeting data, using the CTS-trained MLPs as initial parameter settings. By virtue of this initialization, and the fact that the learning rate was kept small, the MLP would incorporate information from the new target domain (meetings) without "forgetting" the original CTS training data.

Table 2 gives results for various combinations of acoustic models and features, on the two most recent meeting recognition evaluation testsets (RT-04S and RT-05S). The results support a number of interesting observations. We will focus on the results for RT-05S, as that is the larger and more recent testset:

- Adding CTS-trained MLP features, even without adaptation, yields 5.9% relative win (lines c and d).
- Feature adaptation by itself is effective (compare lines a and b), giving about 5.6% relative WER reduction. The gain is about the same as that with Gaussian adaptation alone (lines a and d).
- Gaussian adaptation and feature adaptation are partly additive. Jointly they yield about 8.4% relative WER reduction (lines a and e).
- Comparing an adapted system without MLP features (line c) and a full-adapted MLP system (line e), we find an 8.4% improvement. This is slightly less, but broadly comparable to the 9.9% relative gain obtained on in-domain (CTS) data.

4. CROSS-LANGUAGE EXPERIMENTS

We now address the question of how MLP features behave when applied to other languages. We already had positive results for Mandarin CTS, showing that a *Mandarin-trained* Tandem MLP gives significant improvements [19]. Here, by contrast, we are interested in how much *English-trained* MLPs help the acoustic model for other languages. This seems questionable at first, since each language has its own (sometime radically different) phoneset, and the MLPs are trained for a phone discrimination task that is language dependent. However, languages share phonetic distinctions at the level of articulatory features (such as voicing, frication, and nasality), and to the extent that these distinctions are shared among languages, we can expect the MLP features (trained on any language) to be a useful representation of the acoustic space.

Table 3. Results (CER) without and with MLP features on Mandarin CTS testsets.

	RT-04F devtest		RT-04F eval	
Features	Fast	Full	Fast	Full
Baseline	34.8	32.0	33.1	29.9
w/MLP features	33.5	31.4	31.6	29.0

4.1. Mandarin CTS Experiments

Experiments with Mandarin Chinese we based on our state-of-the-art Mandarin CTS system as used for the NIST RT-04F evaluation [20]. It is similar to the English system (Figure 1) in structure, with minor deviations due to language differences and data availability. For example, the front end omits voicing features, but does include pitch-related features to better capture lexical tone. Most important, the Mandarin phone set contains 65 phone types and encodes lexical tone in the vowels, making it substantially different from the 46-phone set of the English system. Only about 100 hours of acoustic training data were available.

The original Mandarin CTS system included no MLP features. As for English, we retrained the MFCC-based subsystems with feature vectors augmented by Tandem/HATs MLP features. The MLPs in were the same ones used for English. One complication in doing so was that the English MLPs were gender dependent, whereas the Mandarin models were gender independent, due to the small amount of available training data. We simply ran the English CTS gender identification on the Mandarin data, and applied the gender-dependent English MLPs accordingly to each Mandarin speaker. Since both male and female MLPs had been trained to perform the same task (phone classification), we assumed that the resulting features would be compatible across genders, as long as the final dimensionality-reducing PCA transform was the same for both genders. We arbitrarily picked the female PCA transform.

Table 3 summarizes the results, given in character error rates (CER). Adding the MLP features consistently reduces error rates. The reduction is 4.5% relative in the "fast" recognition setup, and 3.0% relative in the full system, measured on the RT-04F evaluation set. The relative improvements are only about 50% to 30% of what we found in the English CTS system, but that is not unexpected given that the features were not trained for the Mandarin task.

4.2. Levantine Arabic CTS

A similar cross-language retraining with English MLP features was done with our RT-04F Levantine Arabic CTS recognition systems [21]. The system structure was again similar to the English system. One difference is that the roles of PLP and MFCC models are swapped, that is, PLP models are used for initial decoding and lattice generation. Accordingly, we added the MLP features to the PLP-based models, and left the MFCC-based models unchanged. Another difference from English is that a nonstandard *factor language model* is used in later stages of the system to better model the complex morphology of Arabic [22]. Again, the acoustic mod-

Table 4. Results (WER) without and with MLP features on Levantine Arabic CTS testsets.

	RT-04F devtest		RT-04F eval	
Features	Fast	Full	Fast	Full
Baseline	46.0	42.1	49.3	46.5
w/MLP features	43.9	40.0	46.7	44.5

els were trained with only a fraction of the amount of data that was available for English: 70 hours. The lack of gender dependence in the Arabic models was handled as described earlier for Mandarin.

Table 4 presents the results. We again see consistent word error reductions with the addition of MLP features. The improvements are between 4.3% and 5.3% relative, about half of what was achieved in English CTS.

5. SUMMARY AND CONCLUSIONS

We have shown that Tandem/HATs MLP features trained on English CTS data can be ported effectively to other domains within the language (meeting speech), and even to other languages (Mandarin and Levantine Arabic). The unchanged feature MLPs, when applied to these new domains, yield about 30% to 60% of the relative improvement (reductions in error rate) as observed for in-domain recognition. This is remarkable especially for cross-language generalization, given the language-dependent, phoneset-specific criterion used in MLP training. MLP feature porting thus represents a novel technique for sharing acoustic training data between languages. This is important because many languages have orders of magnitude less data available than for English.

When porting within the same language, we found that a simple incremental retraining approach was effective for boosting performance of the MLP features in the new domain, yielding almost the full benefit as observed in the CTS domain. A major question for future research will be how to generalize the adaptation of MLPs to new languages with mismatched phonesets.

6. ACKNOWLEDGMENTS

This research was supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811), as well as by DARPA under contract MDA972-02-C-0038 and grant MDA972-02-1-0024. Distribution is unlimited. Barry Chen and Qifeng Zhu trained the original MLP features and gave valuable assistance in feature adaptation. Mari Ostendorf provided valuable guidance in the development of the Mandarin system, and Katrin Kirchhoff contributed to the original Levantine Arabic system.

7. REFERENCES

[1] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *J. Acoust. Soc. Am.*, vol. 87, pp. 1738–1752, Apr. 1990.

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

[3] H. Bourlard and N. Morgan, *Connectionist Speech Recognition. A Hybrid Approach*, Kluwer Academic Publishers, Boston, MA, 1993.

[4] H. Hermansky and S. Sharma, "Temporal patterns (TRAPs) in ASR of noisy speech", *in Proc. ICASSP*, vol. 2, pp. 289–292, Phoenix, AZ, Mar. 1999.

[5] N. Morgan, B. Y. Chen, Q. Zhu, and A. Stolcke, "TRAPping conversational speech: Extending TRAP/Tandem approaches to conversational telephone speech recognition", *in Proc. ICASSP*, vol. 1, pp. 536–539, Montreal, May 2004.

[6] Q. Zhu, A. Stolcke, B. Y. Chen, and N. Morgan, "Using MLP features in SRI's conversational speech recognition system", *in Proc. Interspeech*, pp. 2141–2144, Lisbon, Sep. 2005.

[7] Q. Zhu, B. Chen, N. Morgan, and A. Stolcke, "On using MLP features in LVCSR", in S. H. Kim and D. H. Youn, editors, *Proc. ICSLP*, pp. 921–924, Jeju, Korea, Oct. 2004.

[8] B. Y. Chen, Q. Zhu, and N. Morgan, "Learning long-term temporal features in LVCSR using neural networks", in S. H. Kim and D. H. Youn, editors, *Proc. ICSLP*, Jeju, Korea, Oct. 2004.

[9] H. Misra, H. Bourlard, and V. Tyagi, "New entropy based combination rules in HMM/ANN multi-stream ASR", *in Proc. ICASSP*, vol. 2, pp. 741–744, Hong Kong, Apr. 2003.

[10] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of HMMs", *Computer Speech and Language*, vol. 9, pp. 171–186, 1995.

[11] D. Vergyri, A. Stolcke, V. R. R. Gadde, L. Ferrer, and E. Shriberg, "Prosodic knowledge sources for automatic speech recognition", *in Proc. ICASSP*, vol. 1, pp. 208–211, Hong Kong, Apr. 2003.

[12] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech", *in Proc. ICASSP*, vol. 1, pp. 339–341, Atlanta, May 1996.

[13] N. Kumar, *Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*, PhD thesis, John Hopkins University, Baltimore, 1997.

[14] M. J. Gales, "Maximum likelihood multiple subspace projections for hidden Markov models", *IEEE Trans. Speech Audio Process.*, vol. 10, pp. 37–47, 2002.

[15] H. Jin, S. Matsoukas, R. Schwartz, and F. Kubala, "Fast robust inverse transform SAT and multi-stage adaptation", *in Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, pp. 105–109, Lansdowne, VA, Feb. 1998. Morgan Kaufmann.

[16] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training", *in Proc. ICASSP*, vol. 1, pp. 105–108, Orlando, FL, May 2002.

[17] I. Bulyko, M. Ostendorf, and A. Stolcke, "Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures", in M. Hearst and M. Ostendorf, editors, *Proc. HLT-NAACL*, vol. 2, pp. 7–9, Edmonton, Alberta, Canada, Mar. 2003. Association for Computational Linguistics.

[18] A. Stolcke, X. Anguera, K. Boakye, Ö. Çetin, F. Grézl, A. Janin, A. Mandal, B. Peskin, C. Wooters, and J. Zheng, "Further progress in meeting recognition: The ICSI-SRI Spring 2005 speech-to-text evaluation system", *in Proceedings of the Rich Transcription 2005 Spring Meeting Recognition Evaluation*, pp. 39–50, Edinburgh, July 2005. National Institute of Standards and Technology.

[19] X. Lei, M.-Y. Hwang, and M. Ostendorf, "Incorporating tone-related MLP posteriors in the feature representation for mandarin ASR", *in Proc. Interspeech*, pp. 2981–2984, Lisbon, Sep. 2005.

[20] M. Hwang, X. Lei, T. Ng, M. Ostendorf, A. Stolcke, W. Wang, J. Zheng, and V. Gadde, "Porting Decipher from English to Mandarin", *in Proc. DARPA Rich Transcription Workshop*, Palisades, NY, Nov. 2004.

[21] D. Vergyri, K. Kirchhoff, R. Gadde, A. Stolcke, and J. Zheng, "Development of a conversational telephone speech recognizer for Levantine Arabic", *in Proc. Interspeech*, pp. 1613–1616, Lisbon, Sep. 2005.

[22] D. Vergyri, K. Kirchhoff, K. Duh, and A. Stolcke, "Morphology-based language modeling for Arabic speech recognition", in S. H. Kim and D. H. Youn, editors, *Proc. ICSLP*, pp. 2245–2248, Jeju, Korea, Oct. 2004.