IDIAP RESEARCH REPORT

MULTITASK LEARNING TO IMPROVE ARTICULATORY FEATURE ESTIMATION AND PHONEME RECOGNITION

Ramya Rasipuram Mathew Magimai.-Doss

Idiap-RR-21-2011

JUNE 2011

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

MULTITASK LEARNING TO IMPROVE ARTICULATORY FEATURE ESTIMATION AND PHONEME RECOGNITION

Ramya Rasipuram[1,2] *and Mathew Magimai.-Doss*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale, Lausanne (EPFL), Switzerland
{ramya.rasipuram,mathew}@idiap.ch

ABSTRACT

Speech sounds can be characterized by articulatory features. Articulatory features are typically estimated using a set of multilayer perceptrons (MLPs), i.e., a separate MLP is trained for each articulatory feature. In this report, we investigate multitask learning (MTL) approach for joint estimation of articulatory features with and without phoneme classification as subtask. The effect of number of subtasks in MTL is studied by selecting two different articulatory feature representations. Our studies show that MTL MLP can estimate articulatory features compactly and efficiently by learning the inter-feature dependencies through a common hidden layer representation, irrespective of number of subtasks. Furthermore, adding phoneme as subtask while estimating articulatory features improves both articulatory feature estimation and phoneme recognition. On TIMIT phoneme recognition task, articulatory feature posterior probabilities obtained by MTL MLP achieve a phoneme recognition accuracy of 73.8%, while the phoneme posterior probabilities achieve an accuracy of 74.2%.

Index Terms— multitask learning, articulatory features, posterior probabilities, multilayer perceptrons

1 Introduction

In machine learning and neural networks often it is required to learn a set of multiple related tasks. If the tasks can share what they learn, then learning them together may be better than learning them in isolation. Multitask learning (MTL) is an approach of transfer learning where multiple tasks are learned together and what is learned for each task can help other tasks be learned better [1]. MTL is an inductive transfer mechanism which can be used to improve generalization accuracy, speed of learning and intelligibility of learned models. Multitask learning in neural networks allows features learned at the hidden layer for one task to be useful for other tasks.

In the context of speech processing, MTL has been applied to improve ASR performance (a) in noise by incorporating speech enhancement and gender recognition as additional tasks [2], (b) by high level additional tasks such as gender, broad phoneme classification, grapheme classification [3], (c) on meeting data by jointly learning phone classification and feature mapping from farfield microphone to near field microphone [4]. In addition, multitask neural network has also been used for acoustic-articulatory inversion [5], where the mapping lacks one-to-one relationship between articulation and acoustics.

In this report[1], we investigate the use of MTL framework for joint estimation of articulatory features, such as manner of articulation, place of articulation (Section 2). We study this approach on the TIMIT phoneme recognition task and compare it with the traditional approach of estimating articulatory features using independent classifiers (Section 3). As MTL allows addition of new tasks, we also investigate a framework where both articulatory features and phonemes are learned together. Our studies show that (a) MTL not only yields similar or better system but also a system with fewer number of parameters (about 50% less parameters than independent classifier approach), and (b) adding phoneme classification as an additional task helps in improving both articulatory feature and phoneme recognition (Section 6).

[1]Abridged version of the report is published in ICANN, 2011, pp: 299-306

2 Articulatory Feature Estimation

Phonological studies suggest that each sound unit of a language (phoneme) can be decomposed into a set of features based on the articulators used to produce the sound. Articulatory features define the properties of speech production. There exist different types of articulatory representations of speech, like: binary features, multi-valued features, and government phonological features [6]. In this work, we are interested in multi-valued articulatory features.

2.1 Previous Work

Traditionally, articulatory features are estimated using a set of multilayer perceptron (MLP) classifiers [7, 6, 8], support vector machine classifiers [9], dynamic Bayesian networks (DBNs) [10] etc. Stage: 1 of Figure 1 shows estimation of articulatory features using a set of MLP classifiers. The number of independent MLPs depend upon the way phoneme to articulatory feature maps are derived.

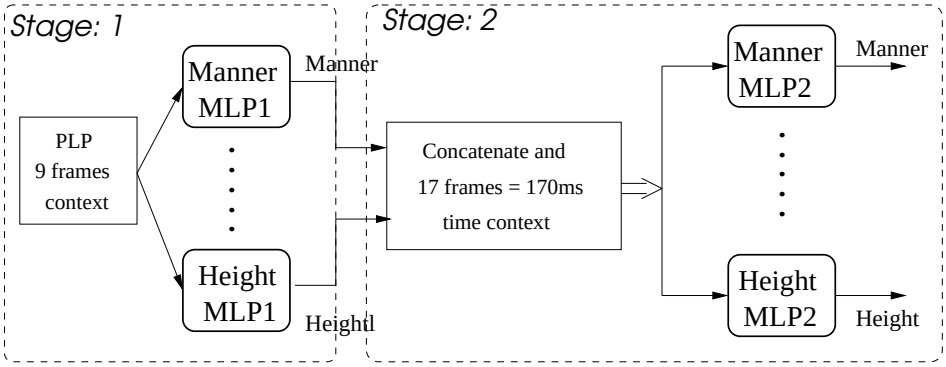

Fig. 1. Hierarchical MLP classifiers for articulatory posterior estimation

In literature, it has been shown that the articulatory feature classification accuracies could be improved by modeling inter-feature dependencies [10]. Along this line, in a more recent work, we showed that by modeling the inter-feature dependencies using a hierarchy of MLP classifiers as shown in Stage: 2 of Figure 1, articulatory feature classification accuracy can be improved, and thereby the phoneme recognition accuracy [8] . The hierarchical approach is originally inspired from [11].

2.2 Proposed Work

In this work, we investigate the use of multitasking MLP (MTL MLP) for joint estimation of articulatory features (as shown in Stage: 1 of Figure 2). The motivation for this is two fold. Firstly, estimating different articulatory features from the same acoustic signal could be considered as a set of interrelated tasks [10]. Traditional, approach of training independent MLPs does not takes it into consideration. Secondly, a system that has fewer number of parameters can be obtained.

Similar to our previous work [8], we also consider a hierarchical approach where a second MTL MLP as shown in Stage: 2 of Figure 2 is trained using the posterior probabilities of articulatory features estimated from Stage: 1 as feature input.

In earlier work, it has been observed that articulatory feature probabilities (articulatory posteriors) and phoneme probabilities (phoneme posteriors) when modeled together can yield better system [8]. Motivated from these observations, we also investigate the importance of phoneme classification as one of the tasks (depicted as dotted line in Figure 2) and examine if MTL could exploit shared hidden layer representation to learn the complementary information.

3 Experimental Setup

In this section we describe the database, phoneme to articulatory feature maps used in the experiments.

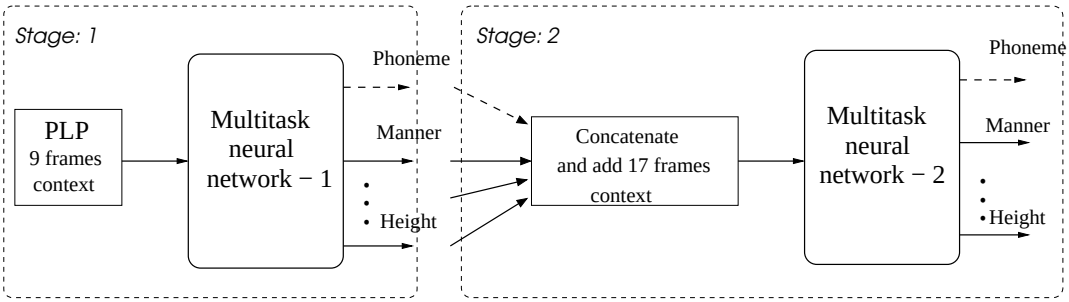

Fig. 2. Hierarchical Multitask MLP classifiers for articulatory (and phoneme) posterior estimation

3.1 Database

TIMIT acoustic-phonetic corpus (excluding the SA sentences) is used in all the experiments. The partitioning of the database as specified in the TIMIT corpus is used. The data consists of 3,000 training utterances from 375 speakers, 696 cross-validation utterances from 87 speakers. Phoneme recognition accuracies are reported on both the complete test set and core set. Complete test set consists of 1344 test utterances from 168 speakers and core set consists of 192 utterances from 24 speakers. The 61 hand labeled phonetic symbols are mapped to set of 39 phonemes with an additional garbage class. The experimental setup is exactly same as the one described in [11].

3.2 Phoneme to Articulatory feature maps

In literature one could find different phoneme to articulatory feature maps [6, 9, 12, 13, 14]. In the case of MTL MLP, one of the difference it brings in is number of tasks that are learned jointly. Therefore, in this work we investigated the phoneme to articulatory feature maps given in John Hopkin's workshop (JHU) [14] and Hosom's thesis [12]. Major differences between the two mappings are: Hosom map is more compact with four features and the cardinality of each of the features is high; JHU map has more features but the cardinality of the features is low compared to Hosom map.

The articulatory features in JHU map consist of manner, place, height, front-back, rounding, glottal state, nasality and vowel (given in Table 2 along with their cardinality).

The articulatory features in Hosom map consist of manner, place and height. Certain changes are made to the mapping defined in Hosom's thesis in order to distinguish all the phoneme classes of the TIMIT. The place class is expanded by adding features like mid-front and mid-back, and the height class by adding features like mid, mid-low, mid-high. Also, vowel articulatory feature is added to the Hosom map (as in JHU map). Table 1 gives the specification of the articulatory features and Table 7 gives the articulatory feature values after the above mentioned modifications on the Hosom's mapping. Glottal state and nasality from JHU are within manner of Hosom map, and frontedness and rounding are within place.

Feature	Cardinality	Feature values
manner	9	approximant, aspirated, flap, fricative, nasal, plosive, voiced fricative, voiced plosive, vowel
place	12	alveolar, dental, dorsal, labial, lateral, retroflex, back, mid-back, mid, front, mid-front, unknown
height	7	low, mid-low, mid, mid-high, high, very-high, max
vowel	20	ae, ah, ao, aw1, aw2, ay1, ay2, eh, er, ey1, ey2, ih, iy, ow1, ow2, oy1, oy2, uh, uw, consonant

Table 1. Values of the articulatory features in Hosom's phoneme to articulatory feature mapping along with their cardinality

4 MLP

The different types of MLP classifiers used in this work are:

1. MTL MLP with articulatory features as tasks (*MTL MLP-af*).
2. MTL MLP with articulatory features and phoneme classification as tasks (*MTL MLP-af+ph*).
3. MLP with one articulatory feature as task, i.e. training a separate classifier for each articulatory feature (*MLP-af*).
4. MLP with phoneme classification as task (*MLP-ph*).

These MLPs can be in the first stage or second stage of hierarchical MLP classifiers as shown earlier in Figures 2 and 1. To compare similar systems, phoneme posteriors are also estimated using a hierarchical MLP classifier described in [11], also shown in Figure 3. All the first stage MLPs use PLP cepstral coefficients with a context window of 9 frames as input and the second stage MLPs use posteriors estimated in the first stage with a temporal context of 17 frames as input.

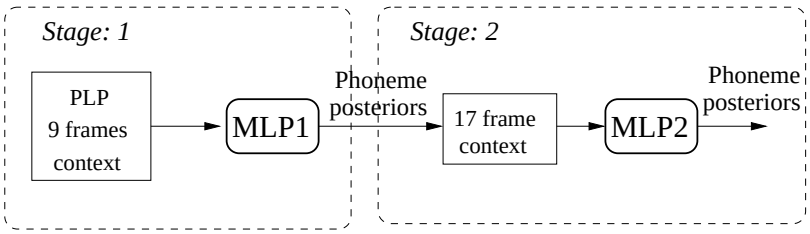

Fig. 3. Hierarchical MLP classifier for phoneme posterior estimation

The hidden layer size of *MTL MLP-af* was optimized on the cross-validation dataset. The same hidden layer size was used for the *MTL MLP-af+ph* and *MLP-ph*. This was done to ascertain the benefit of training jointly both articulatory features and phonemes. As it could be noted that after completion of training *MTL MLP-af+ph* can be split into two MLPs which are of the same size of *MTL MLP-af* and *MLP-ph* (as shown in Figure 4).

In the case of training individual classifiers for each articulatory feature, i.e., *MLP-af*, the size of the hidden layer were determined by fixing the total number of parameters to 35% of the training data following the previous work [8]. The total number of parameters in this system was more than two times of the number of parameters in *MTL MLP-af*.

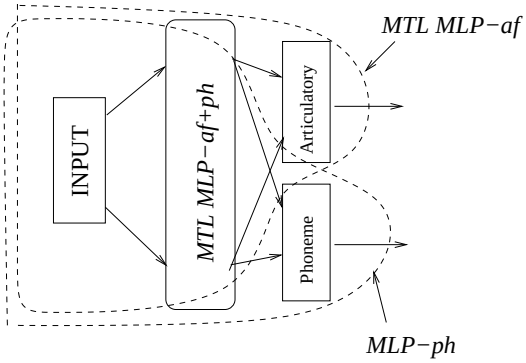

Fig. 4. *MTL MLP-af+ph* estimating phoneme and articulatory posteriors can be split into two MLPs, *MTL MLP-af* and *MLP-ph*

The stopping criterion of the MLPs during training is the cross-validation frame accuracy. All the tasks in the MTL MLP (including the case where phoneme classification is a subtask) are learned with equal learning rate and

equal error weight. It is also observed that the optimal cross-validation performance is obtained for all the articulatory features at the last training epoch.

All the MLPs used in this work are trained using a modified version of ICSI Quicknet software[2] with minimum cross entropy error criterion.

The phoneme recognition experiments were carried out using Kullback-Leibler divergence based hidden Markov model (KL-HMM) system [15]. A brief description about the integration of articulatory feature into KL-HMM system is given in the next section.

5 Integration of AF using KL-HMM acoustic modeling

In KL-HMM acoustic modelling [16], posterior probabilities of sub-word units are directly used as features and the state distribution is parameterized by a reference multinomial distribution (as shown in Figure 5). In [16], the posterior probabilities of phonemes (phoneme posteriors) and in [8], the posterior probabilities of articulatory features (articulatory features) were used as observation features.

In the case of phoneme posteriors the posterior observation feature at time t, $\mathbf{z}_t$ estimated using MLP is given by,

$$\mathbf{z}_t = [z_t^1, \cdots, z_t^D]^T = [P(/aa/|x_t), \cdots, P(/zh/|x_t)]^T \tag{1}$$

where, D is the number of phoneme classes and x_t is input feature given to the MLP. The KL divergence between the multinomial state distribution $\mathbf{y}_i$ and posterior probability feature $\mathbf{z}_t$ is defined as the local matching score for each state, given by,

$$KL(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D} y_i^d \log(\frac{y_i^d}{z_t^d}) \tag{2}$$

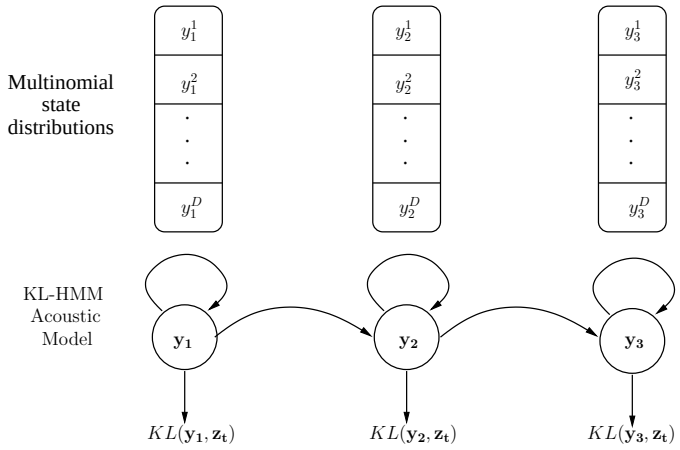

Fig. 5. A three state KL-HMM acoustic model for a phoneme

In the case of articulatory posteriors, the observation feature vector at time t is formed by concatenating the posterior estimates of different articulatory features to a single feature observation vector $\mathbf{z}_t$ as shown below,

$$\mathbf{z}_t = [\mathbf{z}_{t,m}^{D_m}, \ldots, \mathbf{z}_{t,h}^{D_h}]^T, \text{ where,} \tag{3}$$

$$\mathbf{z}_{t,m}^{D_m} = [P(\text{fric}|\mathbf{x}_t), \ldots, P(\text{vowel}|\mathbf{x}_t)]^T$$
$$\mathbf{z}_{t,h}^{D_h} = [P(\text{low}|\mathbf{x}_t), \ldots, P(\text{high}|\mathbf{x}_t)]^T$$

[2] http://www.icsi.berkeley.edu/Speech/qn.html

In this case, the reference multinomial state distribution $\mathbf{y}_i$ is also a stack of multinomial distributions ie,

$$\mathbf{y}_i = [\mathbf{y}_{i,m}^{D_m}, \ldots, \mathbf{y}_{i,h}^{D_h}]^T, \text{ where,} \tag{4}$$

$$\mathbf{y}_{i,m}^{D_m} = [y_{i,m}^1, \ldots, y_{i,m}^{D_m}]^T$$
$$\mathbf{y}_{i,h}^{D_h} = [y_{i,h}^1, \ldots, y_{i,h}^{D_h}]^T$$

where, D_m is the cardinality of the manner class and D_h is the cardinality of the height class. The multinomial state distributions are estimated by optimizing an objective function based on the KL divergence over training dataset [16], given as,

$$KL(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D_m} y_{i,m}^d \log\left(\frac{y_{i,m}^d}{z_{t,m}^d}\right) + \cdots + \sum_{d=1}^{D_h} y_{i,h}^d \log\left(\frac{y_{i,h}^d}{z_{t,h}^d}\right) \tag{5}$$

6 Results

In this section, we first present the results of articulation feature classification studies and then phoneme recognition studies.

6.1 Articulatory feature classification using MTL MLP

Table 2 compares the frame level articulatory feature and phoneme classification accuracies, when estimated using a set of MLPs and MTL MLPs, with JHU phoneme to articulary feature map. The performance of the articulatory features is slightly better when estimated from MTL MLP compared to a set of MLPs. The results also show that along with the articulatory feature classification accuracy, frame level phoneme classification accuracy can also be improved by having phoneme as a subtask in MTL MLP. Similar trends in classification accuracies are observed with Hosom phoneme to articulatory feature map as shown in Table 3.

Task	Cardinality	Chance rates	*MLP-af/MLP-ph*		*MTL MLP-af*		*MTL MLP-af+ph*	
			First stage	Second stage	First stage	Second stage	First stage	Second stage
Manner	8	34.1	86.0	88.1	86.9	88.4	86.9	88.8
Glottal state	5	61.6	92.9	94.5	93.4	94.5	93.4	94.7
Nasality	4	77.9	96.0	96.8	96.4	96.9	96.4	97.0
Place	11	34.1	86.3	88.5	87.0	88.7	87.2	89.3
Height	9	47.7	82.5	85.1	83.8	86.0	83.8	86.5
Frontedness	8	47.7	84.2	86.6	85.3	87.1	85.3	87.6
Rounding	4	67.8	89.9	91.9	91.2	92.9	91.3	93.1
Vowel	22	47.7	81.3	84.5	82.5	84.8	82.7	85.4
Phoneme	40	–	75.1	78.4	–	–	75.6	79.4

Table 2. Frame level articulatory feature and phoneme classification accuracies of individual and MTL MLPs expressed in percentage on the TIMIT cross-validation set with JHU phoneme to articulatory feature map

6.2 Phoneme recognition accuracy

In this section we compare phoneme recognition accuracies of the KL-HMM systems obtained by using phoneme posteriors and articulatory posteriors estimated from MLPs described in Section 4 as feature observations.

6.2.1 First stage results:

Phoneme recognition studies were performed using the posteriors obtained by different first stage of MLPs:

1. *MLP-ph-1*: MLP estimating phoneme posteriors (Stage: 1 in Figure 3).

Task	Cardi-nality	Chance rates	*MLP-af / MLP-ph*		*MTL MLP-af*		*MTL MLP-af+ph*	
			First stage	Second stage	First stage	Second stage	First stage	Second stage
Manner	11	36.8	86.1	88.1	86.7	88.7	86.8	88.9
Place	14	20.0	79.6	82.5	80.1	83.0	80.2	83.4
Height	9	40.1	82.4	85.1	83.7	86.1	83.6	86.3
Vowel	22	47.7	81.3	83.5	82.3	84.9	82.6	85.3
Phoneme	40	–	75.1	78.6	–	–	75.5	79.4

Table 3. Frame level articulatory feature classification accuracies of individual and Multitask MLPs expressed in percentage on the TIMIT cross-validation set with Hosom phoneme to articulatory feature map

2. *MLP-af-1*: a set of MLPs estimating articulatory posteriors (Stage: 1 in Figure 1).
3. *MTL MLP-af-1*: MTL MLP estimating articulatory posteriors without phoneme subtask (Stage: 1 in Figure 2).
4. *MTL MLP-af+ph-1*: MTL MLP estimating articulatory posteriors and phoneme posteriors i.e., MTL MLP with phoneme as one of the subtask (Stage: 1 in Figure 2).

Table 4 presents the phoneme recognition accuracies of the above systems on the test set of TIMIT database with two phoneme to articulatory feature maps, JHU and Hosom. Results show that the phoneme recognition accuracy obtained using articulatory posteriors estimated from MTL MLP is significantly better than the system using posteriors from independent MLPs in case of JHU map, and slightly better in case of Hosom map. The addition of phoneme subtask to the MTL MLP further improves the accuracy of the system using articulatory posteriors as well as the system using phoneme posteriors (especially with JHU map).

MLP	MLP hidden units	MLP o/p units		Posteriors used	Accuracy	
		JHU	Hosom		JHU	Hosom
MLP-ph-1	3500	40		phoneme	70.2 (69.2)	
MLP-af-1	Not applicable	71	56	articulatory	67.4 (66.4)	68.4 (67.5)
MTL MLP-af-1	3500	71	56	articulatory	68.9 (67.8)	68.7 (67.8)
MTL MLP-af+ph-1	3500	111	96	articulatory phoneme	69.2 (68.5) 70.4 (69.3)	69.5 (68.6) 70.0 (69.0)

Table 4. Phoneme recognition accuracy expressed in percentage on the TIMIT test set (core set), using phoneme posteriors and articulatory posteriors as features in KL-HMM

6.2.2 Second stage results:

In this section we present the phoneme recognition accuracies of different systems obtained by using the posteriors from different second stage of MLPs whose input is the output of the first stage of MLPs. The baseline second stage of MLPs used for comparison are:

1. *MLP-af-2*: a set of hierarchical MLPs estimating articulatory posteriors (Stage: 2 in Figure 1).
2. *MLP-ph-2*: hierarchical MLP estimating phoneme posteriors (Stage: 2 in Figure 3).

Two different posteriors can be obtained from hierarchical MTL MLP systems based on the presence or absence of phoneme subtask:

1. *MTL MLP-af-2*: hierarchical MTL MLP estimating articulatory posteriors i.e., MTL MLP without phoneme subtask.
2. *MTL MLP-af+ph-2*: hierarchical MTL MLP estimating articulatory posteriors and phoneme posteriors i.e., MTL MLP with phoneme as one of the subtask.

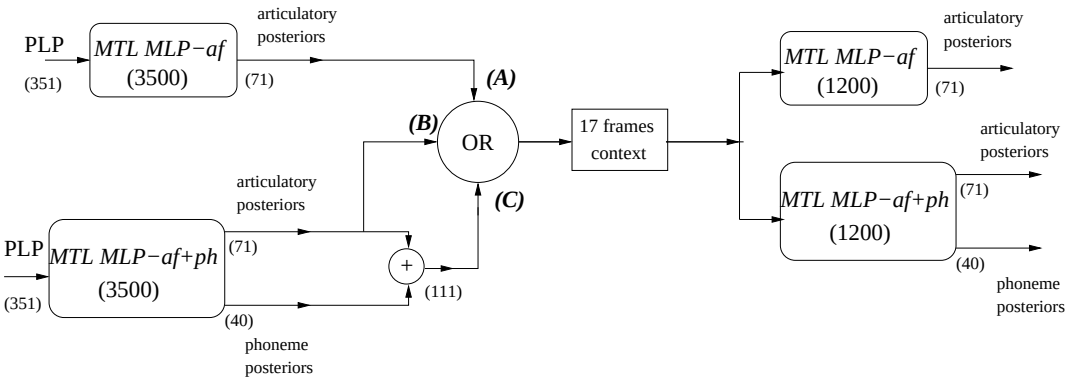

Fig. 6. Hierarchical MTL MLP systems with inputs, hidden and outputs specified

However, the input to the above MTL MLPs can be the articulatory posteriors obtained from two first stage MLPs *MTL MLP-af-1* or *MTL MLP-af+ph-1* (shown as inputs ***(A)*** and ***(B)*** in Figure 6 respectively). Also, a hierarchical MTL MLP system was built where the input of the second MTL MLP consisted of phoneme posteriors and articulatory posteriors (shown as input ***(C)*** in Figure 6).

Table 5 presents the phoneme recognition accuracies obtained by using baseline MLP and MTL MLP posteriors as features in KL-HMM system. The results show that performance of the system using articulatory posteriors from MTL MLP without phoneme task is comparable to the system using articulatory posteriors from a set of MLPs. The second stage MTL MLP with input as articulatory posteriors from *MTL MLP-af+ph* (discarding phoneme posteriors) further improves the performance slightly. Thus, indicating that articulatory features could be learned better when phoneme classification is also a subtask. Overall, in all the cases MTL MLP with phoneme as subtask improves the phoneme recognition performance of system that uses articulatory posteriors as well as system that uses phoneme posteriors.

Hierarchical MLP		MLP i/p units	MLP o/p units	Posteriors used	Phoneme accuracy
First stage	Second stage				
MLP-ph-1	*MLP-ph-2*	680	40	phoneme	73.0 (72.1)
MLP-af-1	*MLP-af-2*	1207	71	articulatory	72.0 (71.2)
MTL MLP-af-1 (Input ***(A)*** in Fig 6)	*MTL MLP-af-2*	1207	71	articulatory	72.2 (71.3)
	MTL MLP-af+ph-2		111	articulatory	72.3 (71.5)
				phoneme	72.7 (72.2)
articulatory posteriors of *MTL MLP-af+ph-1* (Input ***(B)*** in Fig 6)	*MTL MLP-af-2*	1207	71	articulatory	72.4 (71.9)
	MTL MLP-af+ph-2		111	articulatory	72.5 (71.7)
				phoneme	73.3 (72.8)
MTL MLP-af+ph-1 (Input ***(C)*** in Fig 6)	*MTL MLP-af-2*	1887	71	articulatory	72.7 (72.3)
	MTL MLP-af+ph-2		111	articulatory	**73.2 (72.6)**
				phoneme	**74.0 (73.4)**

Table 5. Phoneme recognition accuracy expressed in percentage on the TIMIT test set (core set), using phoneme posteriors and articulatory posteriors as features in KL-HMM. Articulatory features are obtained using JHU map. Also given in the table number of input and output units of MLPs

The MTL MLP with phoneme as subtask at both the stages gave the best performance of 73.2% for articulatory posteriors and 74.0% for phoneme posteriors. It is important to note that the system benefited from both phoneme input and MTL of articulatory and phoneme tasks.

Furthermore, to examine if articulatory features are indeed contributing for the improvement obtained using phoneme posteriors, the phoneme posteriors from first stage MTL MLP (*MTL MLP-af+ph*) are used as input to a second stage phoneme MLP (*MTL MLP-ph*). The accuracy of the system using these phoneme posteriors is 72.6%

(compared to 74.0% using phoneme posteriors from *MTL MLP-af+ph*).

Table 6 presents similar results for Hosom map. The results show that irrespective of number of sub-tasks MTL MLP improves over a set of MLPs. However, Hosom map performs slightly better than JHU map. This may be due to (a) compactness of articulatory features or (b) local score calculation as Hosom map results in sum of fewer KL-divergences in Equation 5.

Hierarchical MLP		MLP i/p units	MLP o/p units	Posteriors used	Phoneme accuracy
First stage	Second stage				
MLP-ph-1	*MLP-ph-2*	680	40	phoneme	73.0 (72.1)
MLP-af-1	*MLP-af-2*	952	56	articulatory	72.1 (71.6)
MTL MLP-af-1 (Input ***(A)*** in Fig 6)	*MTL MLP-af-2*	952	56	articulatory	72.6 (71.9)
	MTL MLP-af+ph-2		96	articulatory	72.9 (72.4)
				phoneme	73.5 (72.6)
articulatory posteriors of *MTL MLP-af+ph-1* (Input ***(B)*** in Fig 6)	*MTL MLP-af-2*	952	56	articulatory	72.8 (72.4)
	MTL MLP-af+ph-2		96	articulatory	73.1 (72.7)
				phoneme	73.6 (72.9)
MTL MLP-af+ph-1 (Input ***(C)*** in Fig 6)	*MTL MLP-af-2*	1887	56	articulatory	73.6 (73.0)
	MTL MLP-af+ph-2		96	articulatory	**73.8 (73.3)**
				phoneme	**74.2 (73.7)**

Table 6. Phoneme recognition accuracy expressed in percentage on the TIMIT test set (core set), using phoneme posteriors and articulatory posteriors as features in KL-HMM. Articulatory features are obtained using Hosom map. Also given in the table number of input and output units of MLPs

7 Discussion and Conclusions

Our studies show that MTL provides a framework for efficient and compact estimation of articulatory posteriors compared to a set of MLPs. Furthermore, jointly training articulatory features and phoneme improves both articulatory feature classification and phoneme recognition. We hypothesize that *MTL MLP-af+ph* through a shared hidden layer learns to exploit the complementary information present in phoneme and articulatory tasks. This is partly supported by the fact that we do not achieve significant improvement in phoneme recognition accuracy (74.1% compared to 74.0% with phoneme posteriors alone) when concatenating phoneme posteriors and articulatory feature posteriors (as done in our previous work [8]).

The results comparing the two phoneme to articulatory feature maps (JHU and Hosom) showed that irrespective of number of sub-tasks, similar trends in terms of articulatory feature classification and phoneme recognition accuracy are observed. Furthermore, in preliminary (ongoing) ASR studies we have observed trends similar to phoneme recognition at word recognition level.

In this work, during the training of MTL MLPs all the tasks were given equal importance. It may be interesting to study the effect of giving one or a few tasks more importance than others. Our future work will also focus on addition of more subtasks, such as gender, rate-of-speech estimation, and performing full-fledged ASR studies.

Acknowledgments

This work was supported by the Swiss NSF through the grants "Flexible Grapheme-Based Automatic Speech Recognition (FlexASR) " and the National Center of Competence in Research (NCCR) on "Interactive Multimodal Information Management" (www.im2.ch). The authors would like to thank Joe Frankel, CSTR, Edinburgh for fruitful discussions as well as providing the multitasking MLP software.

8 References

[1] Rich Caruana, "Multitask Learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[2] S. Parveen and P. Green, "Multitask Learning in Connectionist Robust ASR using Recurrent Neural Networks," in *Proceedings of EUROSPEECH*, 2003, pp. 1813–1816.

[3] J. Stadermann, W. Koska, and G. Rigoll, "Multi-task Learning Strategies for a Recurrent Neural Net in a Hybrid Tied-Posteriors Acoustic Model," in *Proc. of Interspeech*, 2005, pp. 2993–2996.

[4] J. Frankel, Ö. Çetin, and N. Morgan, "Transfer Learning for Tandem ASR Feature Extraction," in *Proceedings of MLMI*, 2007, pp. 227–236.

[5] K. Richmond, "A Multitask Learning Perspective on Acoustic-Articulatory Inversion," in *Proc. of Interspeech*, 2007.

[6] S. King and P. Taylor, "Detection of Phonological Features in Continuous Speech using Neural Networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 333–353, 2000.

[7] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, pp. 303–319, 2002.

[8] R. Rasipuram and M. Magimai.-Doss, "Integrating Articulatory Features using Kullback-Leibler Divergence based Acoustic Model for Phoneme Recognition," in *Proc. of ICASSP*, 2011, pp. 5192– 5195.

[9] O. Scharenborg, V. Wan, and R.K. Moore, "Towards Capturing Fine Phonetic Variation in Speech using Articulatory Features," *Speech Communication*, vol. 49, pp. 811–826, 2007.

[10] J. Frankel, M. Wester, and S. King, "Articulatory feature recognition using dynamic Bayesian networks," in *Computer Speech & Language*, 2007, vol. 21(4), pp. 620–640.

[11] J. Pinto, G. Sivaram, M Magimai.-Doss, H. Hermansky, and H. Bourlard, "Analysis of MLP based Hierarchical Phoneme Posterior Probability Estimator," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 225–241, 2011.

[12] J.-P Hosom, "Automatic Phoneme Alignment Based on Acoustic-Phonetic Modeling," in *Proc. of ICSLP*, 2002, vol. I, pp. 357–360.

[13] K. Livescu et al., "Articulatory Feature-based Methods for Acoustic and Audio-Visual Speech Recognition: 2006 JHU Summer Workshop Final Report," `http://www.clsp.jhu.edu/ws2006/groups/afsr/documents/WS06AFSR_final_report.pdf`, 2008.

[14] J. Frankel, M. Magimai-Doss, S. King, K. Livescu, and Ö. Çetin, "Articulatory Feature Classifiers Trained on 2000 hours of Telephone Speech," in *Proc. of Interspeech*, 2007.

[15] G. Aradilla, J. Vepa, and H. Bourlard, "An Acoustic Model Based on Kullback-Leibler Divergence for Posterior Features," in *Proc. of ICASSP*, 2007, pp. 657–660.

[16] G. Aradilla, H. Bourlard, and M. Magimai.-Doss, "Using KL-Based Acoustic Models in a Large Vocabulary Recognition Task ," in *Proc. of Interspeech*, 2008, pp. 928–931.

Timit phoneme	Manner	Place	Height	Vowel
sil	silence	silence	silence	silence
ae	vowel	mid-front	low	ae
ah	vowel	mid	mid	ah
ao	vowel	back	mid-low	ao
aw1	vowel	mid-front	low	aw1
aw2	vowel	mid-back	high	aw2
ay1	vowel	back	low	ay1
ay2	vowel	mid-front	high	ay2
b	voiced stop	labial	max	consonant
ch	stop	front	max	consonant
dh	voiced fricative	dental	max	consonant
d	voiced stop	alveolar	max	consonant
dx	flap	alveolar	max	consonant
eh	vowel	mid-front	mid	eh
er	vowel	mid	mid	er
ey1	vowel	front	mid-high	ey1
ey2	vowel	mid-front	high	ey2
f	fricative	labial	max	consonant
g	voiced stop	dorsal	max	consonant
hh	aspirated	unknown	max	consonant
ih	vowel	mid-front	high	ih
iy	vowel	front	very-high	iy
jh	voiced stop	front	max	consonant
k	stop	dorsal	max	consonant
l	approximant	lateral	very-high	consonant
m	nasal	labial	max	consonant
ng	nasal	dorsal	max	consonant
n	nasal	alveolar	max	consonant
ow1	vowel	back	mid	ow1
ow2	vowel	mid-back	high	ow2
oy1	vowel	back	mid-low	oy1
oy2	vowel	mid-front	high	oy2
p	stop	labial	max	consonant
r	approximant	retroflex	mid-low	consonant
s	fricative	alveolar	max	consonant
sh	fricative	front	max	consonant
th	fricative	dental	max	consonant
t	stop	alveolar	max	consonant
uh	vowel	mid-back	high	uh
uw	vowel	back	very-high	uw
v	voiced fricative	labial	max	consonant
w	approximant	back	very-high	consonant
y	approximant	front	very-high	consonant
z	voiced fricative	alveolar	max	consonant
oth	reject	reject	reject	reject

Table 7. Values for each of articulatory features in Hosom's phoneme to articulatory feature mapping along with their cardinality