Articulatory Feature Classifiers Trained on 2000 hours of Telephone Speech

Joe Frankel[1,2], Mathew Magimai-Doss[2], Simon King[1], Karen Livescu[3], Özgür Çetin[2].

[1]University of Edinburgh, [2]ICSI, [3]MIT

joe@cstr.ed.ac.uk

Abstract

This paper is intended to advertise the public availability of the articulatory feature (AF) classification multi-layer perceptrons (MLPs) which were used in the Johns Hopkins 2006 summer workshop. We describe the design choices, data preparation, AF label generation, and the training of MLPs for feature classification on close to 2000 hours of telephone speech. In addition, we present some analysis of the MLPs in terms of classification accuracy and confusions along with a brief summary of the results obtained during the workshop using the MLPs. We invite interested parties to make use of these MLPs.

1. Introduction

Steady interest persists among the speech recognition research community as to how speech production knowledge may be used to improve word error rates (WERs) [1]. Given that a significant portion of the variation present in spontaneous, conversational speech can be expressed in terms of articulatory strategy, many researchers have argued that better results could be achieved by *modelling* the underlying processes of co-articulation and assimilation, rather than simply *describing* their effects on the speech signal.

A set of discrete multi-level articulatory features (AFs) is one way in which speech production information may be represented [2, 3, 4]. Table 1 gives the feature specification used in this work, along with the cardinality of each group.

group	cardinality	feature values
place	10	alveolar, dental, labial, labio-dental, lateral, none, post-alveolar, rhotic, velar
degree	6	approximant, closure, flap, fricative, vowel
nasality	3	- +
rounding		- +
glottal state	4	aspirated, voiceless, voiced
vowel	23	aa, ae, ah, ao, aw1, aw2, ax, ay1, ay2, eh, er, ey1, ey2, ih, iy, ow1, ow2, oy1, oy2, uh, uw, nil
height	8	high, low, mid, mid-high, mid-low, very-high, nil
frontness	7	back, front, mid, mid-back, mid-front, nil

Table 1: *Specification of the multi-valued articulatory features used in this work. Note that each group also has 'silence' and 'reject' classes.*

AF recognition is a task which has been considered by a number of researchers (including the authors of this paper), as being one method by which speech production knowledge may be inferred and incorporated into a larger system whilst staying firmly within a data-driven statistical approach.

Kirchhoff showed that within the hybrid artificial neural network (ANN) / hidden Markov model (HMM) paradigm, error rates could be reduced using AF-based classifiers [3]. More recently, this summer's Johns Hopkins (JHU) workshop [5] gave promising results using feature-based classifiers within the tandem HMM [6] paradigm, which represents the current state-of-the-art in speech recognition.

Multi-layer perceptrons (MLPs) have been successfully used for AF classification. One of the difficulties in training such an MLP is generating the feature values which are used as targets. These are typically created by mapping from time-aligned phone labels, which can prove to be a bottleneck in scaling systems based on corpora such as TIMIT and OGI Numbers, which include manual time-aligned phone transcripts, up to larger tasks.

The experience gained at ICSI was that larger MLPs, trained on a greater amount of data, lead to improvements in phone-based tandem automatic speech recognition (ASR). This suggests that similar benefits could be gained by training AF classification MLPs on larger corpora. In this paper we describe articulatory feature classifier MLPs trained on close to 2000 hours of spontaneous telephone quality speech from the Fisher and Switchboard corpora.

We were fortunate in being able to use the compute and disk resources at ICSI to train the MLPs, and further had access to phone-level alignments produced by Decipher, SRI's large vocabulary speech recognition system. We consider that a number of researchers may be interested in using such a resource, and therefore based the trainings and associated tasks (such as acoustic parameterization) on freely available tools. The MLP weights along with other resources are available from www.cstr.ed.ac.uk/research/projects/featureMLPs.

2. Data preparation

The MLPs were trained on data drawn from the Fisher and Switchboard2 corpora. No data from Switchboard1 was used to ensure that none of the utterances in the SVitchboard (small vocabulary Switchboard set) [7], which was the planned subject of experimentation in the JHU workshop, were present in the MLP training data. Conversation sides were pre-segmented into utterances, and randomly assigned to either the training or cross validation (CV) sets, with 10% of conversation sides being used for cross validation. This gave train and CV sets of 1776 and 225 hours of data respectively.

HTK was used to parameterize the acoustic waveforms as 12 PLP cepstra plus energy calculated every 10ms within 25ms Hamming windows. The base features were then mean and variance normalized on a per-speaker basis, and then 1st and

2nd order derivatives appended to give a 39-dimensional feature vector. The HTK configuration files and global variance against which the speaker normalizations are calculated are given on the MLP web-page.

3. Feature mappings

The feature labels used as targets in training the MLPs were mapped from time-aligned phone labels. The phone to feature mapping is given in the Appendix in Table 6, and was designed based on the phone set in the first column (we refer to this as the WS06 phone set). SRI provided alignments of Fisher and Switchboard data which necessitated a mapping from the SRI to WS06 phone-sets. The main difference was that the SRI set does not separate stops into closure and release portions, and additionally diphthongs are considered to be a single unit.

In order to allow flexibility in how stops and diphthongs were treated, the mappings were made from SRI states (3 per phone) to WS06 phones. After an examination of a set of example alignments, the following conventions were chosen: diphthongs were split at the central frame of the middle state, and for stops the first two states were assigned to be the closure portion, and the last state was set to be release.

All feature groups include a silence class, and a reject label which is assigned (by the SRI aligner) to frames which align poorly with the transcript. In addition to the classes in Table 1, an extra symbol which appears in Table 6 is '&'. This is used where a feature should take the value of the first right-context non-'&' symbol, for example, [hh] doesn't have an inherent rounding value but instead takes the rounding value of the following phone.

4. MLP configuration

Following the tandem ASR [6] approach, a context window of 9 frames (central frame plus 4 frames each of left and right context) was used on the input layer for all MLPs. Given the 39 dimensional input feature, this amounts to 351 input units. The number of units on the output layer of each MLP corresponds to the cardinality of the feature group. The numbers of hidden units are given in Table 2, and were set according to the following rationale. Setting 2400 hidden units corresponds to a ratio of training frames to parameters of 1000:1. This was taken as an upper limit and used for the highest cardinality group (vowel), and half of this, 1200, for the lowest cardinality groups (nasal, rounding). The others are spaced in the interval [1200, 2400] in proportion to the log of the number of output units.

feature group	cardinality	hidden units
place	10	1900
degree	6	1600
nasality	3	1200
rounding	3	1200
glottal state	4	1400
vowel	23	2400
height	8	1800
frontness	7	1700

Table 2: *Number of hidden units used by each AF MLP.*

Sigmoid activation functions were used at the hidden layer, and a softmax on the outputs. The initial learn rate was set to 0.0004, and then followed a scheme in which training continues with fixed learn rate until the CV accuracy increases by less than 0.5%, after which the learn rate is halved at each epoch. Training is considered complete when the CV accuracy increase between epochs again falls below 0.5%.

The `Quicknet` tools developed at ICSI were used for training the MLPs. These are freely available for download at `www.icsi.berkeley.edu/Speech/qn.html`.

5. Cross validation

Table 3 shows the framewise accuracy for each MLP for both training and cross-validation sets, along with chance rates based on the class priors. The accuracy is measured against the phone-derived feature labels, with the effect that some errors may in fact be feature changes which we would wish to capture, though are not expressed in the alignments. In addition, confusion matrices were estimated, and can be found on the MLP web-page mentioned above. We make the following observations on the

feature group	Train	CV	chance
place	76.5%	76.2%	33.6%
degree	78.0%	77.8%	34.3%
nasal	90.7%	90.5%	47.1%
rounded	87.9%	87.7%	42.6%
glottal state	87.3%	87.1%	42.3%
vowel	73.6%	73.3%	33.3%
height	75.7%	75.4%	34.3%
frontness	76.1%	75.8%	34.2%

Table 3: *Framewise AF classification accuracy for each MLP, on training and cross-validation data.*

MLP trainings for each feature group:

Place The class "none" is used during vowels. Across all other classes, 14% of frames are misclassified as "none". The classification accuracies for classes "dental" and "labio-dental" are below 50%, and are mainly misclassified as "alveolar".

Degree Classes "approximant", "fricative", "closure", and "flap" are usually misclassified as vowel. The class "flap" has the lowest classification accuracy of 35% and, is most frequently confused with "closure", "fricative" or "vowel".

Nasal Out of the 3 classes "sil", "-" and "+", "-" has the highest recognition accuracy of 95%. Most confusions for the values "sil" and "+" are with "-".

Rounded Similar trend to the nasal feature group.

Glottal state The "voiced" class has the highest classification accuracy. Classes "sil", "aspirant" and "voiceless" are most often confused with "voiced" class, in particular there are many misclassifications of "aspirant".

Vowel The classification accuracy for the classes vary between 17% and 65% other than those for the classes "sil" and "nil", which is assigned to consonants. The majority of vowel misclassifications are with the "nil" class.

Height The classification accuracy for the classes vary between 40% and 67% other than for the classes "sil" and "nil". As with the vowel feature group, the height classes tend to be misclassified as "nil".

Frontness The pattern of errors is similar to that of the height feature group.

The two main trends across all feature groups are firstly that silence is recognized with high accuracy ($\approx$ 85%), and secondly that misclassified classes are usually recognized as the speech class which has the highest prior probability.

6. ASR Results

During the JHU 2006 workshop, the articulatory feature MLP outputs were used in a number of ways. The first was for observation probability estimation in hybrid AF-based ASR, though performance did not match that of a standard phone-based HMM baseline. However, embedded training in which the MLPs were adapted on SVitchboard (SVB) data led to a significant reduction in the difference between the systems [8].

Another approach which was investigated was to use the AF MLP classifiers within a tandem approach [6]. This work is reported in [5], and we present a summary of the key findings below.

Tandem ASR involves the use of MLPs to provide non-linear transforms of the feature space. The benefits arise from the extra contextual information provided by the input window, and from providing a mapping into a space in which class separation is maximized. Typically, the classes are phones, though using the AF MLPs means that the mapping is to a feature-based space.

Once computed, the MLP posteriors are subject to a log transformation and dimensionality reduction via principal components analysis (PCA). They are then appended to standard acoustic features for use in an HMM system.

Features (monophone HMMs)	WER (%)
PLP baseline	67.7
PLP + phone tandem, SVB-trained	63.0
PLP + AF tandem, SVB-trained	62.3
PLP + AF tandem, Fisher-trained	59.7

Table 4: Word error rates for various monophone systems on the Svitchboard 500-word E set.

Table 4 shows word error rates (WER) for a number of monophone systems on a Svitchboard 500-word vocabulary task. The introduction of tandem features is shown to give a substantial decrease in WER over a PLP-only baseline, from 67.7% to 63.0%. The next pair of results compare tandem MLPs trained against phone and AF targets on Svitchboard, which provides close to 4 hours of training material. The AF-based tandem system gives a slightly lower WER than the phone-based tandem, 62.3% compared with 63.0%. The final row shows the monophone results using the 2000-hour AF MLPs described in this paper. The WER is further reduced to 59.7%.

Features (triphone HMMs)	WER (%)
PLP baseline	59.2
PLP + AF tandem, Fisher-trained	55.0

Table 5: Word error rates for various triphone system on the Svitchboard 500-word E set.

Table 5 shows the performance of triphone systems on the same Svitchboard task. The addition of tandem features generated with the 2000-hour AF MLPs gives a significant reduction in word error rate compared with a PLP-only baseline, from 59.2% to 55.0%. These results demonstrate one method by which AF-based MLPs may be usefully employed in an ASR system.

7. Summary

This paper has presented a survey of the design and training of a set of articulatory feature classification MLPs. Further information and MLP weights are available from the project web-page given in Section 1. In addition, instructions for invocation of the Quicknet forward pass routine `qnsfwd` (generates MLP posteriors) are given, along with the resources needed to compute compatible and suitably scaled front-end features using HTK.

Finally, an exploration of manual labelling of articulatory features and comparison with the classification made by the MLPs described in this paper can be found in [9].

8. References

[1] Simon King, Joe Frankel, Karen Livescu, Erik McDermott, Korin Richmond, and Mirjam Wester, "Speech production knowledge in automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, February 2007.

[2] E. Eide, J.R. Rohlicek, H Gish, and S. Mitter, "A linguistic feature representation of the speech waveform," in *Proc. ICASSP-93*, 1993, pp. 483–486.

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

[4] J. Frankel and S. King, "A hybrid ANN/DBN approach to articulatory feature recognition," in *Proceedings of Eurospeech*, Lisbon, Portugal, 2005, CD-ROM.

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

[6] D. Ellis, R. Singh, and S. Sivadas, "Tandem acoustic modeling in large-vocabulary recognition.," in *Proc. ICASSP*, 2001.

[7] S. King, C. Bartels, and J. Bilmes, "SVitchboard 1: Small vocabulary tasks from Switchboard 1.," in *Proc. Interspeech*, Lisbon, Portugal, 2005.

[8] K. Livescu et al., "Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU Summer Workshop," in *Proc. ICASSP*, Honolulu, April 2007.

[9] K. Livescu et al., "Manual transcription of conversational speech at the articulatory feature level," in *Proc. ICASSP*, Honolulu, April 2007.

WS06 phone	place	degree	nasality	rounding	glottal state	vowel	height	frontness
sil	silence	silence	silence	silence	silence	silence	silence	silence
aa	none	vowel	-	-	voiced	aa	low	back
ae	none	vowel	-	-	voiced	ae	low	mid-front
ah	none	vowel	-	-	voiced	ah	mid	mid
ao	none	vowel	-	+	voiced	ao	mid-low	back
aw1	none	vowel	-	-	voiced	aw1	low	mid-front
aw2	none	vowel	-	+	voiced	aw2	high	mid-back
ax	none	vowel	-	-	voiced	ax	mid	mid
ay1	none	vowel	-	-	voiced	ay1	low	back
ay2	none	vowel	-	-	voiced	ay2	high	mid-front
bcl	labial	closure	-	-	voiced	nil	nil	nil
b	labial	fricative	-	-	voiced	nil	nil	nil
tcl	alveolar	closure	-	-	voiceless	nil	nil	nil
ch	post-alveolar	fricative	-	&	voiceless	nil	nil	nil
dcl	alveolar	closure	-	-	voiced	nil	nil	nil
d	alveolar	fricative	-	-	voiced	nil	nil	nil
dh	dental	fricative	-	-	voiced	nil	nil	nil
dx	alveolar	flap	-	-	voiced	nil	nil	nil
eh	none	vowel	-	-	voiced	eh	mid	mid-front
er	rhotic	approximant	-	&	voiced	er	mid	mid
ey1	none	vowel	-	-	voiced	ey1	mid-high	frt
ey2	none	vowel	-	-	voiced	ey2	high	mid-front
f	labio-dental	fricative	-	-	voiceless	nil	nil	nil
gcl	velar	closure	-	-	voiced	nil	nil	nil
g	velar	fricative	-	-	voiced	nil	nil	nil
hh	none	vowel	-	&	aspirated	&	&	&
ih	none	vowel	-	-	voiced	ih	high	mid-front
iy	none	vowel	-	-	voiced	iy	very-high	frt
dcl	alveolar	closure	-	-	voiced	nil	nil	nil
jh	post-alveolar	fricative	-	&	voiced	nil	nil	nil
kcl	velar	closure	-	-	voiceless	nil	nil	nil
k	velar	fricative	-	-	voiceless	nil	nil	nil
l	lateral	closure	-	-	voiced	nil	nil	nil
m	labial	closure	+	-	voiced	nil	nil	nil
n	alveolar	closure	+	-	voiced	nil	nil	nil
ng	velar	closure	+	-	voiced	nil	nil	nil
ow1	none	vowel	-	+	voiced	ow1	mid	back
ow2	none	vowel	-	+	voiced	ow2	high	mid-back
oy1	none	vowel	-	+	voiced	oy1	mid-low	back
oy2	none	vowel	-	-	voiced	oy2	high	mid-front
pcl	labial	closure	-	-	voiceless	nil	nil	nil
p	labial	fricative	-	-	voiceless	nil	nil	nil
r	rhotic	approximant	-	&	voiced	nil	nil	nil
s	alveolar	fricative	-	-	voiceless	nil	nil	nil
sh	post-alveolar	fricative	-	&	voiceless	nil	nil	nil
tcl	alveolar	closure	-	-	voiceless	nil	nil	nil
t	alveolar	fricative	-	-	voiceless	nil	nil	nil
th	dental	fricative	-	-	voiceless	nil	nil	nil
uh	none	vowel	-	+	voiced	uh	high	mid-back
uw	none	vowel	-	+	voiced	uw	very-high	back
v	labio-dental	fricative	-	-	voiced	nil	nil	nil
w	labial	approximant	-	+	voiced	nil	nil	nil
y	post-alveolar	approximant	-	-	voiced	nil	nil	nil
z	alveolar	fricative	-	-	voiced	nil	nil	nil
zh	post-alveolar	fricative	-	&	voiced	nil	nil	nil
-	reject	reject	reject	reject	reject	reject	reject	reject

Table 6: *Mapping used to generate feature targets from time-aligned phone labelling. The phone set is given in the left column. The symbol '&' denotes that the value of the first non-& right-context phone is used.*