Phone Classification using a Non-Linear Manifold with Broad Phone Class Dependent DNNs

Linxue Bai, Peter Jančovič, Martin Russell, Philip Weber, Steve Houghton
School of Engineering, University of Birmingham

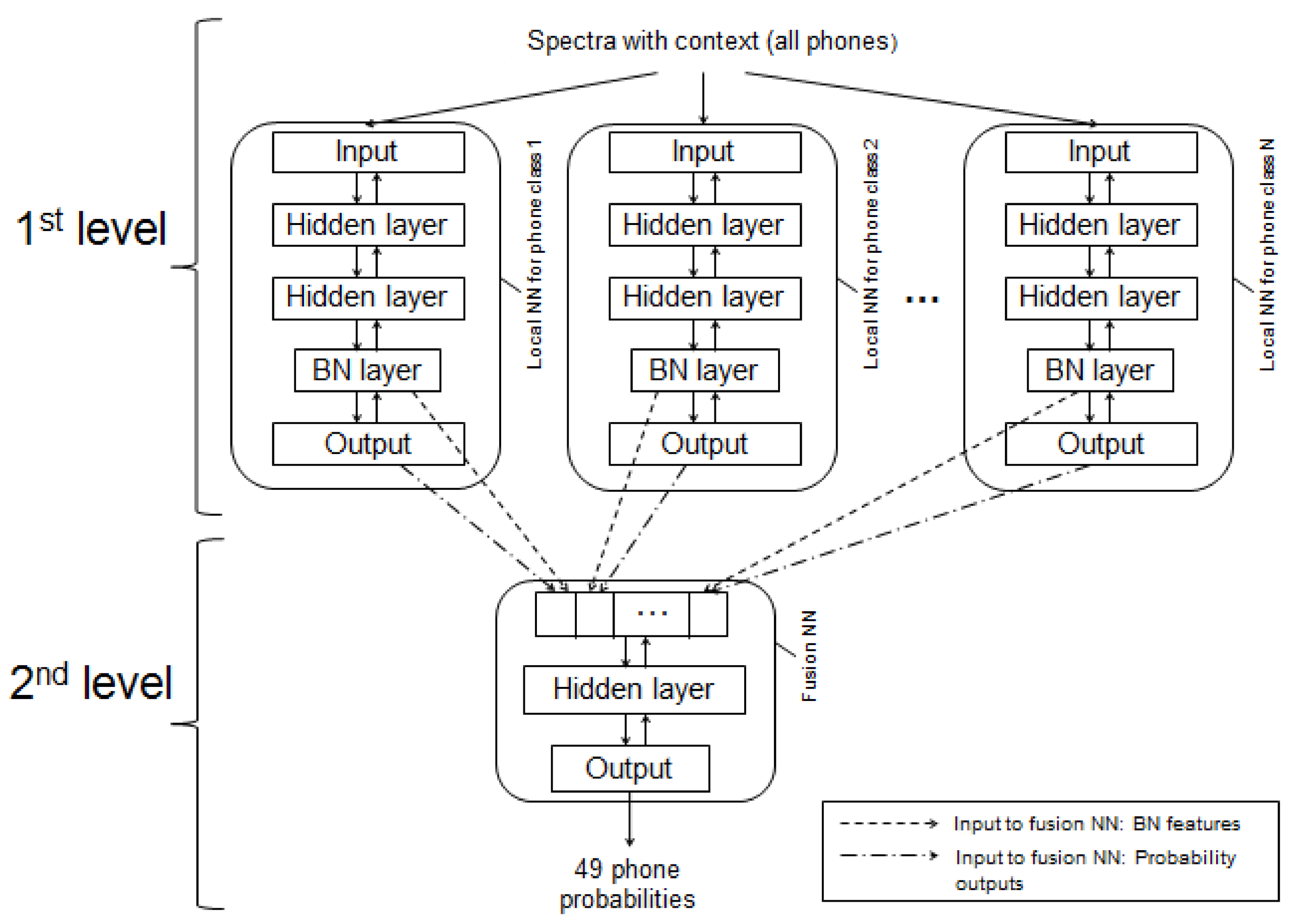
UNIVERSITY OF BIRMINGHAM

Introduction

- The objective is to determine whether it is advantageous for phone-classification of feature vectors to treat the acoustic space as a non-linear manifold, in which several broad phone class (BPC)-dependent DNNs rather than a single DNN are used.
- This extends our previous study of very low dimensional bottleneck features (BNFs) [1,2], and the work by Huang et.al [3] on a manifold structure learning linear mappings.

Experimental Setup

System structure



- 1st level: A set of parallel BPC-dependent DNNs - Every local DNN is trained with every frame by using an additional node for “out-of-the-class” phones.
- 2nd level: A fusion network that makes the final phone classification decision - The NN input is BN features or probability outputs from the 1st level.

Phonetic broad classes [3]

Table 1: *Phonetic broad classes used to define the set of local DNN-based projections.*

Group	Phonetic class	Phone label
Q_1	Plosive	/g/, /d/, /b/, /k/, /t/, /p/
Q_2	Strong fricative	/s/, /z/, /sh/, /zh/, /ch/, /jh/
Q_3	Weak fricative	/f/, /v/, /th/, /dh/, /hh/
Q_4	Nasal/Flap	/m/, /n/, /en/, /ng/, /dx/
Q_5	Semi-vowel	/l/, /el/, /r/, /w/, /y/
Q_6	Short vowel	/ih/, /ix/, /ae/, /ah/, /ax/, /eh/, /uh/, /aa/
Q_7	Long vowel	/iy/, /uw/, /ao/, /er/, /ey/, /ay/, /oy/, /aw/, /ow/
Q_8	Silence	/sil/, /epi/, /q/, /vcl/, /cl/
Q_9	$Q_5 \cup Q_6 \cup Q_7$:	Semi-vowel, Short vowel, Long vowel
Q_{10}	$Q_1 \cup Q_3$:	Plosive, weak fricative
Q_{11}	$Q_5 \cup Q_6$:	Semi vowel, Short vowel
Q_{12}	$Q_5 \cup Q_7$:	Semi vowel, Long vowel
Q_{13}	$Q_6 \cup Q_7$:	Short vowel, Long vowel
Q_{14}	$Q_1 \cup Q_2 \cup \ldots \cup Q_8$:	All phones

BPCs used to train local DNNs

Table 2: *The sets $D_1, \ldots, D_5$ of BPCs used to train local BPC-dependent DNNs in the two-level system.*

Broad phone class	Experimental setup D_1	D_2	D_3	D_4	D_5
$Q_1 - Q_8$	X	X	X	X	X
Q_9		X	X		
Q_{10}			X	X	X
Q_{11}				X	X
Q_{12}				X	X
Q_{13}				X	X
Q_{14}					X
# of local DNNs	8	9	10	12	13

- Data: TIMIT (incl. labels and time-stamp information)
- Local DNN Input: 26-dim Mel filterbanks with context of ± 5 frames.
- DNN Training: Deep belief networks (DBN) with GRBM/RMB pretraining and stochastic gradient descent using Theano.
- Evaluation: (i) on all frames in the core test set, (ii) on only centre frames of phone segments (also need to finetune DNN).

Experimental Results - Phone Classification Performance

Table 3: *Phone classification accuracy obtained using all signal frames and using only the centre frames of each phone.*

	All frames		Centre frames	
Global DNN	67.60 (*avg*)		76.81 (*avg*)	
	69.05 (*avg*+3*std*)		77.58 (*avg*+3*std*)	
Local DNNs	Fusion net input		Fusion net input	
	Softmax	BNF	Softmax	BNF
D_1 (*avg*)	69.05*	68.78	77.45	77.03
D_2 (*avg*)	69.44*	69.23*	77.85	77.75
D_3 (*avg*)	69.56*	69.24*	78.31*	78.11*
D_4 (*avg*)	69.76*	69.31*	78.59*	78.08
D_5 (*avg*)	70.01*	69.63*	78.93*	78.70*

(“*” indicates a pass of McNamar's significant test in more than 95% pairwise comparisons; In all the experiments we tried to keep the number of parameters the same.)

Conclusions

- The BPC-dependent DNNs provided small but significant improvements in phone classification accuracy in comparison to a single global DNN.
- It is advantageous to also include local DNNs focusing on a combination of some BPCs.
- The use of the softmax outputs as input to the fusion network provided slightly better results than the bottleneck outputs.

Experimental Results - Bottleneck feature visualisation

Linear Discriminant Analysis (LDA) was applied to visualise the BNFs from the global and the local DNNs. The local DNN for plosive is used below as an example.

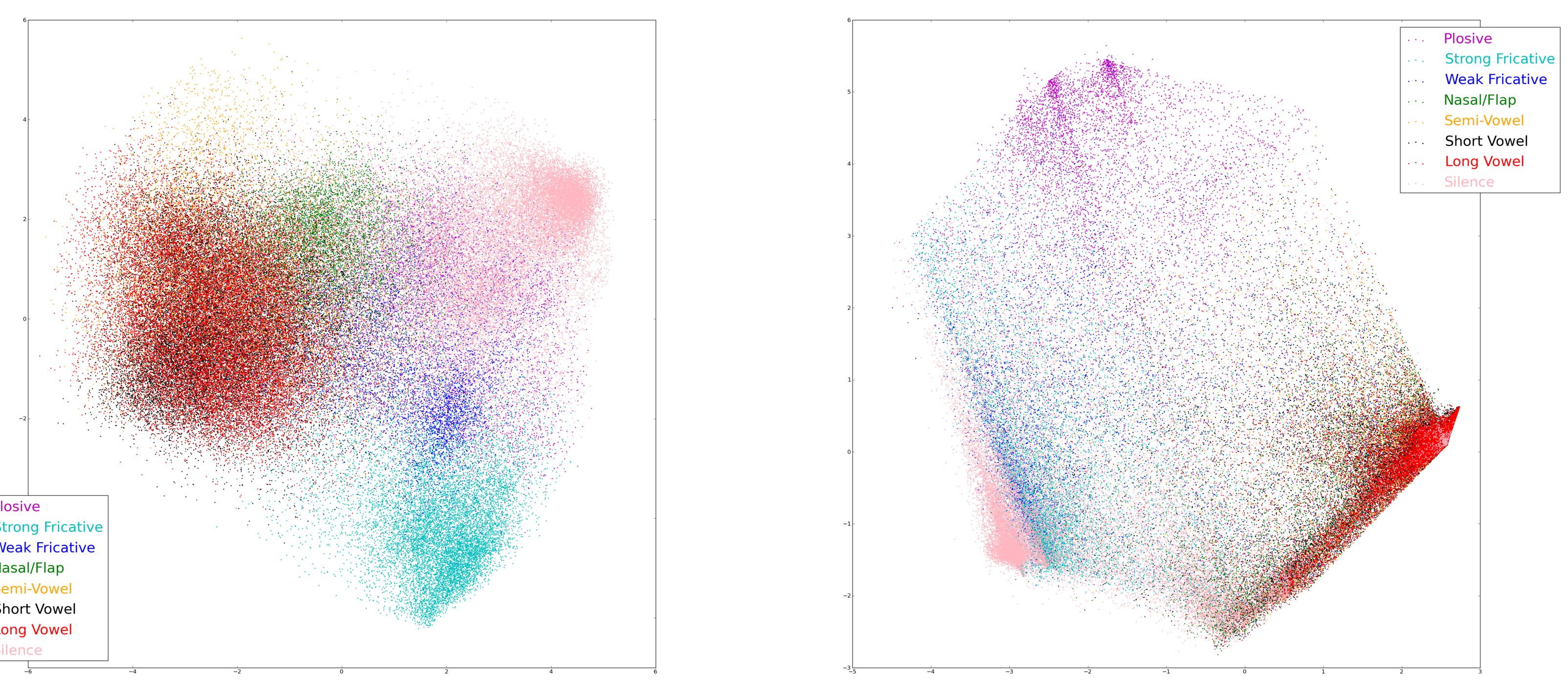


Fig 1: 1st vs 2nd LDA of BNFs (all phones) from a global DNN (left) and a local DNN for plosives (right)

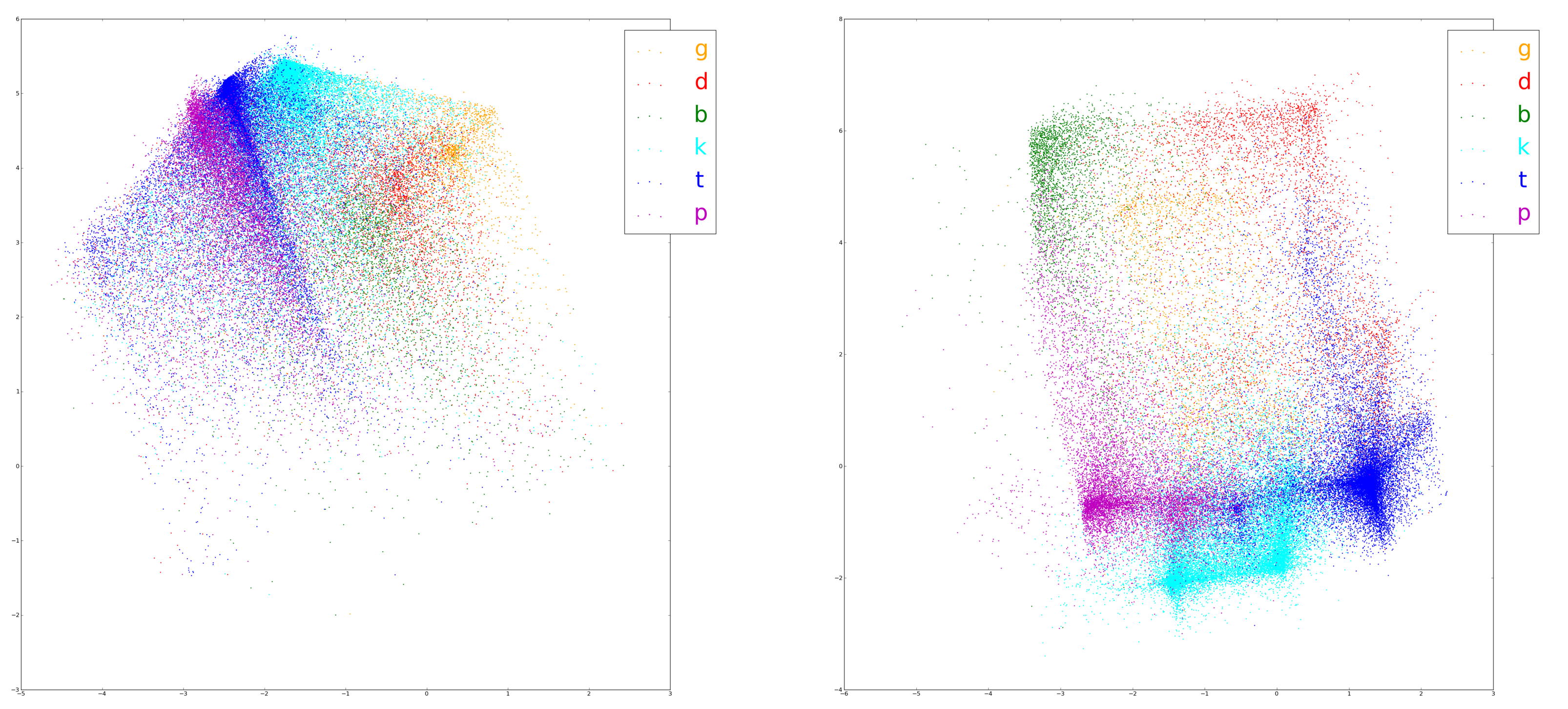


Fig 2: Visualisations of BNFs (plosive phones) from a local DNN for plosives: 1st vs 2nd LDA (left) and 3rd vs 4th LDA (right)

Conclusion

- Local DNNs learn clearer local structures, which may be related to speech production mechanisms.

References

1. L. Bai, P. Jančovič, M. Russell, and P. Weber, *“Analysis of a low dimensional bottleneck neural network representation of speech for modelling speech dynamics”*, *Proc. Interspeech 2015*, pp. 583–587.
2. P. Weber, L. Bai, M. Russell, P. Jančovič, and S. Houghton, *“Interpretation of low dimensional neural network bottleneck features in terms of human perception and production”*, *Proc. Interspeech 2016*, pp. 3384–3388.
3. H. Huang, Y. Liu, L. ten Bosch, B. Cranena, and L. Boves, *“Locally learning heterogeneous manifolds for phonetic classification”*, *Computer Speech and Language*, pp. 28–45, 2016.