A Neural Architecture for Computing Acoustic-phonetic Invariants

Elaine Tsiang
et@monowave.com

Abstract

The proposed neural architecture consists of an analytic lower net, and a synthetic upper net. This paper focuses on the upper net. The lower net performs a 2D multiresolution wavelet decomposition of an initial spectral representation to yield a multichannel representation of local frequency modulations at multiple scales. From this representation, the upper net synthesizes increasingly complex features, resulting in a set of acoustic observables at the top layer with multiscale context dependence. The upper net also provides for invariance under frequency shifts, dilatations in tone intervals and time intervals, by building these transformations into the architecture. Application of this architecture to the recognition of gross and fine phonetic categories from continuous speech of diverse speakers shows that it provides high accuracy and strong generalization from modest amounts of training data.

1. INTRODUCTION

Neural networks for continuous speech recognition have been predominantly used in hybrid systems[1][2]. Segmentation of continuous speech is an incidental consequence of model selection in the HMM approach. The apparent elimination of the problem of segmentation comes at a cost - the search space increases exponentially with the perplexity of the grammar. If segmentation can be based more on acoustic observables than is the case in current HMM systems, then it may be decoupled from sentence model selection, and the search space made independent of perplexity. Such observables would be portable across task domains, and to a significant degree, across languages. Towards the computation of such suitable observables, we have proposed a neural model of acoustic perception[3]. The complete model consists of a cochlear filter bank for initial spectral analysis[4], an analyzing lower net, and a synthesizing upper net.

Segmentation has been traditionally based on phonemes, the acoustic correlates of which seem to depend on contexts larger than a segment. Phonemes are therefore local representations of relatively nonlocal information. The elementary tonotopic features (ETFs) computed by the lower net[3][5], spanning various tone intervals, and time intervals, are also such local representations of nonlocal information. Recent neurophysiological data also indicate there exist cortical structures for such representations[6].

It seems logical, therefore, to define phonemes, or some closely-related observables, as functions of ETFs. By observables, we mean functions of input signal only, independent of any pre-compiled knowledge. The upper net is a means for realizing such functional relations. Human recognition of continuous speech is further characterized by robustness against certain transformations arising out of the differences among speakers, or speaking styles, even from the same speaker. Frequency shifts are induced by differing vocal chord lengths, or differing vocalizations (e.g. falsetto). Tone interval dilatations result from varying formant frequencies and intervals, as a result of differing vocal tract dimensions, or accents. Time dilatations result from differing speaking rates. Robustness may be defined as invariance under limited ranges of these transformations. The acoustic-phonetic observables will have similar invariance if we build such transformations into the architecture.

2. STRUCTURAL ELEMENTS

The basic structural element for both the lower and upper nets is the *tonotopic feature (TF).* A *TF* is a family of self-similar 2-D filters over tonotopy-time. The self-similarity in tonotopy endows the features with frequency shift invariance. We refer to any 2-D representation in tonotopy and time as a *map*. A *layer* is a logical grouping of maps. The set of TFs that operate on the maps of a layer, and output to maps of another layer is referred to as a *tract*.

$$map_{c[i]}(x_0, t_0) = \sum_t \sum_x f_{c[i] \leftarrow d[j]}(x - x_0, t - t_0) map_{d[j]}(x, t) \tag{1}$$

The subscript $c[i]$ indicates that the map is in layer c. $c[i] \leftarrow d[j]$ indicates that the TF inputs from map $d[j]$ and outputs to map $c[i]$. The TFs have limited extents in x, t. We refer to its 2-D matrix of coefficients, $f(\Delta x, \Delta t)$, as its receptive field. $map_{c[i]}(x_0, t_0)$ may be interpreted as a measure of similarity between the TFs receptive field and the input pattern on the patch of channels and frames around channel x_0 and time t_0. The TFs in a tract have the same receptive field sizes and decimations. The maps of a given layer all have the same number of channels. A map may receive outputs from more than one TF, which may be summed, thresholded and compressed:

$$map_{c[i]}(x_0, t_0) = g\left(\left(\sum_{j=m}^{n} \sum_t \sum_x f_{c[i] \leftarrow d[j]}(x - x_0, t - t_0) map_{d[j]}(x, t)\right) - \Theta_{c[i]}\right) \tag{2}$$

where $\Theta_{c[i]}$ is the threshold weight. In Eq. (2), the input maps are all in the same layer. In general, they may belong to different layers.

Viewing the entire network as one single equivalent tract from the input layer to the final output layer, the cumulative effect of the TFs is to give the filters in this equivalent tract large receptive fields. Figure 4 shows an example of afferently decreasing chan-

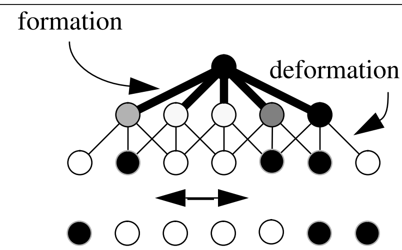

Figure 1. Local Shifts as Dilatations

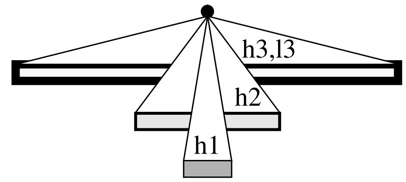

Figure 2. Nesting Contexts

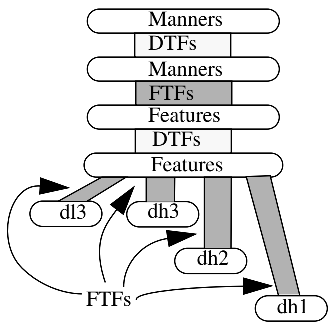

Figure 3. Manner Upper Net

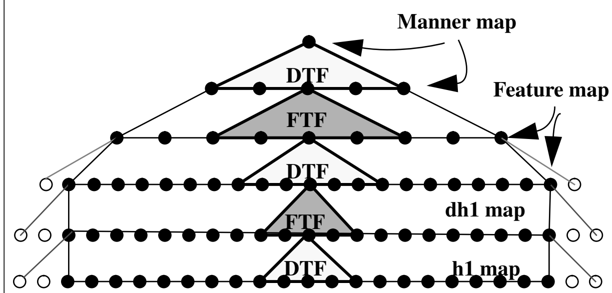

Figure 4. Map and Receptive Field sizes

nels and efferently expanding receptive fields in 5 layers. The upper net and the lower net combined typically have 10 layers, resulting in extensive context dependence, at multiple scales, jointly in tonotopy and time.

3. THE LOWER NET

An *elementary tonotopic feature (ETF)* is a TF that detects the presence of *tones*, their *onset, rise or fall*, corresponding to 4 basic degrees of frequency modulation. A multiresolution ETF is a family of ETFs obtained by dilatations in both tonotopy and time. For the reported experiments, we used 4 levels of resolution, each coarser level obtained by a 2x dilatation in tonotopy and time. [5] describes an implementation of a partially orthogonal Gabor transform-based wavelet decomposition, which breaks the spectral content into 32 frequency modulation components - 8 highpass maps at 2x dilatation (layer h1), 8 highpass maps at 4x dilatation (layer h2), and 8 highpass and 8 lowpass maps at 8x dilatation (layers h3 and l3).

4. THE UPPER NET

Two types of layers, *formation* and *deformation*, alternate in the upper net. A deformation tract inputs one or more formation layers and outputs to a deformation layer. The *deformation TFs (DTFs)* provide limited local shifts[7] in tonotopy and time and sum and compress the outputs from one or more DTFs, providing feature alternation in the case of multiple DTFs. The amount of compression depends on the DTF coefficients. With sufficient compression, dilatating an interval leaves the DTF output within the matching tolerance of the next formation TF(Figure 1). The DTF coefficients also determine the range of dilatation. Because the coefficients are adapted in training, the network learns the tolerable dilatations from data. Specificly, the output layers from the lower net are first subjected to deformation tracts which output to the corresponding deformation layers dh1,dh2,dh3 and dl3.

A formation tract inputs one or more deformation layers and outputs to a formation layer. The *formation TFs (FTFs)* of such tracts are similar to the ETFs, but the filters match to more complex patterns. Each complex tonotopic feature, when trained, matches to a pattern at the h1 scale, nested within a pattern at the h2 scale, nested within a pattern at the h3 and l3 scale(Figure 2). The pattern at each level of resolution provides the context for all finer levels of resolution, in both time and tonotopy.

The dual formation-deformation layering of this upper architecture is similar to the Neocognitron[8], minus the structures for normalization. The computations are the same as a standard multilayer perceptron. Also unlike the Neocognitron, all filter coefficients, including the DTFs', and thresholds, are adapted through training on speech data.

5. GROSS-CATEGORY RECOGNITION

"Manner of articulation" is loosely defined as the categorization of speech sounds according to the type of stricture and other articulatory attributes except place[9]. Table 1 gives the categories we use. Figure 3 gives the specific architecture. Figure 4 gives the map sizes and receptive field sizes in both tonotopy and time. There are 12 feature maps, requiring a total of 7225 weights. Because of the broad categories, this task forces the network to discover features common to apparently different phonemes. Table 2 shows that the network has indeed discovered some well-known phonological features: nasality (more accurately, "anti-nasality"), obstruency (stops and fricatives) and voicing. Feature 7 suggests that fricatives and nasals also have closures, i.e., constrictions.

Speech Data

In Table 3, test set 3 was recorded with a Olympus Pearlcorder L200 microcassette recorder, and a SONY MTL F-96 microphone, in an office environment

Performance

The most significant result in Table 4 is the consistency of the per-speaker average and standard deviation across test set 1(same

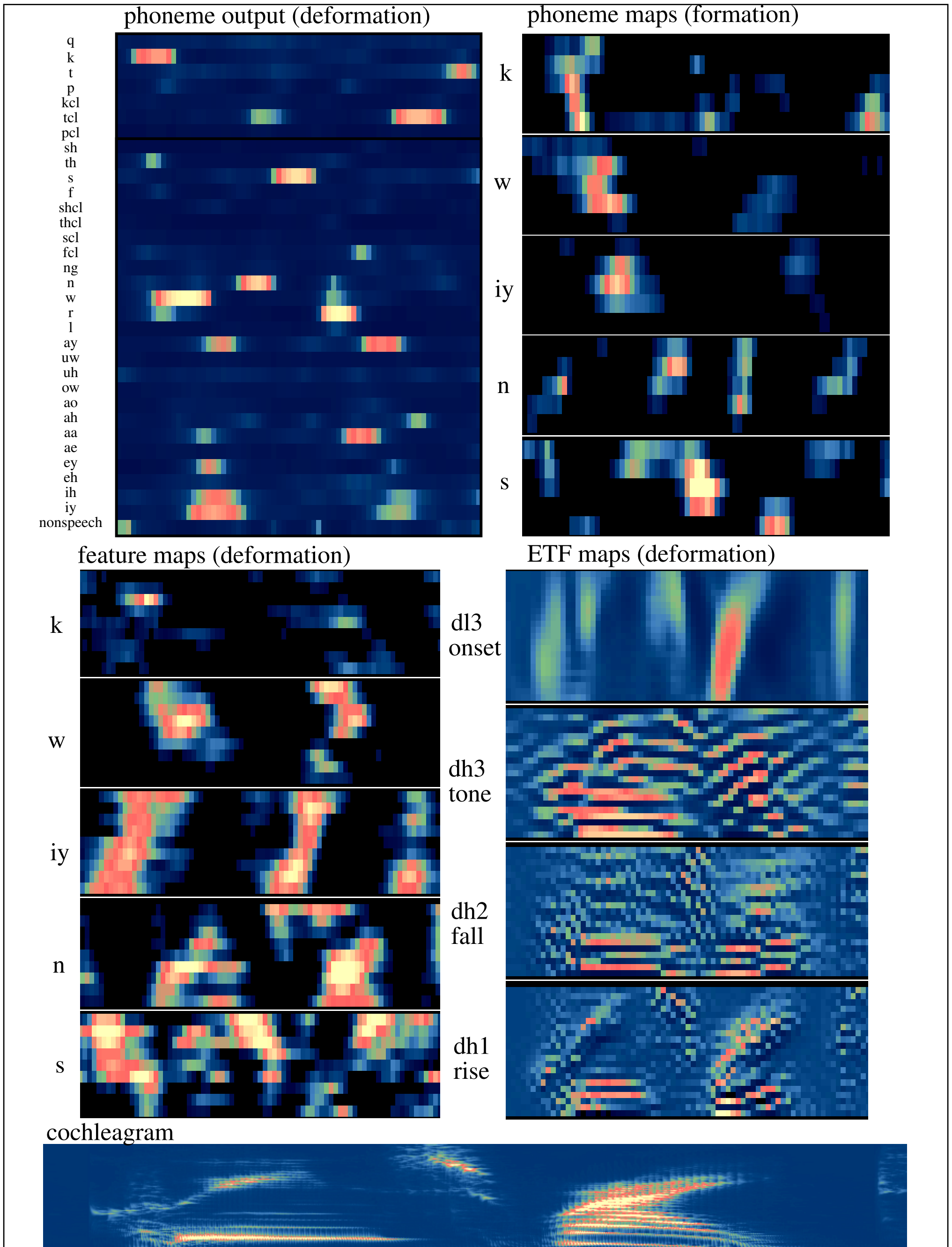

Figure 5. Maps for "queen's right"

speakers, different sentences, same recording environment), test set 2(different speakers, different sentences, same recording environment) and test set 3 (different speech databases, different recording environments). This shows that the network generalizes well across different speakers, recording environments, and task domains. Note that for both experiments, the results are strictly acoustic-phonetic, unaided by language-model-based segmentation.

6. PHONEME RECOGNITION

We built this recognizer for a voice-controlled chess-playing demonstrator. The vocabulary consists of 25 words for playing chess which require 33 phonemes, omitting 1 glide(y), 1 nasal(m) and h, with no distinction between voiced and unvoiced consonants; but a group of fricative closures is added. There are 7 independent subnets, one for each major group of phonemes, for a total of 43046 weights. Each subnet has an architecture similar to the manner recognizer. Figure 5 gives a sample of the maps and the final outputs for a short utterance.

Speech Data

We collected the samples in an office environment using a Nady

TABLE 1. Manners

Manner	Symbol	TIMIT phonemes
nonspeech	0	pau, h#
voiceless stop releases	P	p,t,k,ch
voiceless stop closures	C	pcl,tcl,kcl
voiced stop releases	B	b,d,dx,g,jh
voiced stop closures	Q	bcl,dcl,gcl,q
voiceless fricatives	F	s,sh,f,th,hh
voiced fricatives	V	z,zh,v,dh,hv
nasals	M	m,n,nx,ng
liquids	R	l,r
glides	W	y,w
vowels	A	iy,ih,eh,ey,ae,aa,aw, ay,ah,ao,oy,ow,uh,uw, ux,er,ax,ix,axr, ax-h,em,en,eng,el

TABLE 2. Manner-feature connections

	features											
	0	1	2	3	4	5	6	7	8	9	10	11
P	x	x	x	x	x		x		x	x		
C	x	x	x	x	x		x	x	x	x		
B	x	x	x	x	x	x	x		x	x		
Q	x	x	x	x	x	x	x	x	x	x		
F	x	x	x	x	x			x	x	x		
V	x	x	x	x	x		x	x	x	x		
M		x	x			x	x	x				
R	x	x	x	x	x						x	
W	x	x	x	x	x							x
A	x	x	x	x			x					

TABLE 3. Data sets for manner recognition

set	#utterances	#sentences	#speakers	recording
train	144	118	36(18male 18female)	TIMIT
test 1	216	172	same	TIMIT
test 2	80	80	16(8m,8f)	TIMIT
test 3	80	80	27	cassette
test 4	80	80	23	NTIMIT

TABLE 4. Manner accuracy

set	# frames	per-frame	per-speaker	
			average	std
Train	28360	83.7%	83.4%	4.9%
Test 1	41708	72.5%	72.3%	3.7%
Test 2	16947	71.6%	71.6%	3.1%
Test 3	13644	70.2%	70.1%	4.9%
Test 4	15189	61.5%	61.5%	4.6%

MCM-400 headset microphone and a Radio Shack SSM-100 mixer. The training set consists of 910 utterances of 88 chess commands from 84 male and 62 female speakers. The test set consists of 492 utterances of the same commands from 33 male and 20 female speakers.

Performance

In addition to Table 5, the network also performed well on informal tests with out-of-vocabulary phrases containing the syllabic onsets and rhymes present in the chess commands, suggesting that it has developed features important in the transitions from one phoneme to the next.

TABLE 5. Phoneme accuracy

	# frames	per-frame
training set	130086	92.4%
test set	66170	87.2%

7. CONCLUSION

As with other purely neural network-based systems, the experiments require precisely hand-labelled data, which has restricted the scope of the speech data we used, and makes benchmark comparisons with HMM or hybrid approaches such as [10] difficult. However, in the class of neural networks incorporating significant contextual information, including TDNNs[11] and Gamma MLPs[12], this paper gives the first significant results on continuous speaker-independent tasks.

Current efforts explore alternative observables that are more directly related to transitions, and their utilization in finite state machines for word recognition.

References

[1] N. Morgan and H. Bourlard, “Continuous Speech Recognition - an introduction to the hybrid HMM/connectionist approach”, *IEEE Signal Processing Magazine, v*ol. 12, no. 3, pp. 25-42, May 1995.

[2] S. Young, “A Review of Large-vocabulary Continuous-speech Recognition”, *IEEE Signal Processing Magazine,* vol. 13, no.5, pp. 45-54, Sep., 1996.

[3] E. Tsiang, “System for recognizing speech”, U.S. Patent 5377302, 1994.

[4] E. Tsiang, “A Cochlea Filter Bank for Speech Analysis”, *Proc. International Conference on Signal Processing Applications and Technology*, pp.1674-1678, 1997.

[5] E. Tsiang, “Multiresolution Elementary Tonotopic Features for Speech Perception”, *Proc. International Conference on Neural Networks, pp. 575-579,* June 1997.

[6] C.E. Schreiner and M.L. Sutter, “Functional topography of cat primary auditory cortex: distribution of integrated excitation”, *J. Neurophysiology*, vol. 64, no. 5, pp. 1442-1459, 1990.

[7] E. Barnard and D. Casasent, “Shift Invariance and the Neocognitron”, *Neural Networks,* vol. 3, no. 4, pp. 403-410, 1990.

[8] K. Fukushima, “Neocognitron: A Hierarchical Neural Network Capable of Visual Pattern Recognition”, *Neural Networks,* vol. 1, no. 2, pp. 119-130, 1988.

[9] J. Laver, *Principles of Phonetics*, Cambridge University Press, 1994.

[10] A. Robinson, “An application of recurrent nets to phone probability estimation”, IEEE Trans. Neural Networks, vol. 5, no. 2, pp. 299-304, 1994.

[11] K.J. Lang, A.H. Waibel and G.E. Hinton, “A Time-Delay Neural Network Architecture for Isolated Word Recognition”, Neural Networks, vol. 3, no. 1, pp. 23-43, 1990.

[12] S. Lawrence, A.C. Tsoi, A.D. Back, “The Gamma MLP for Speech Phoneme Recognition”, Neural Information Processing 8, MIT Press, 1996.