# Articulatory Feature Extraction Using Temporal Flow Neural Networks[*]

Shawn Chang[†]<shawnc@cs.berkeley.edu>

10 May 1998

**Abstract**

A number of interesting properties of articulatory features motivate its incorporation into speech recognition systems. This work explores the possibility of extracting articulatory features from pre-processed speech signals (modulation spectrogram) using a special temporal flow model (TFM) neural network. Good performance was obtained with the TFMs on each feature category, and the results were comparable to or better than that of the multilayer perceptron (MLP) with much larger number of parameters on the same tasks.

We also show results of applying the TFM outputs in the estimation of posterior phoneme probabilities with another MLP. Reasonable phoneme frame accuracy were obtained, and the TFM/MLP system gave complimentary resultls to that of an MLP-only system. Combination of the results from the TFM/MLP system and MLP-only system gained significant improvements in the phoneme frame accuracy.

Analysis on the trained TFM network gave interesting node activation correlation patterns, corresponding to useful cues for articulatory feature extraction. We also discuss the effect of recurrent links and mutiple time-delayed links of TFM on the context window size.

## 1 Introduction

The motor theory of perception suggests that human brains interpret the received speech signals in terms of the neural pattern production that are needed to articulate the same incoming speech. If this theory were to stand, it must be true that there are enough information in the speech signals to recover the articulatory features (AF), which can be any suitable physical description of the vocal tract during speech production.

No matter the motor theory were true or not, there are several potential advantages of incorporating AFs into speech processing systems. The slow varing nature of our speech aparatus motivates an AF-based model, especially for the continuous speech processing, where coarticulation is abundant. The physical limitations of the vocal tract determine that both anticipitory and carry-over coarticulation can occur in different articulators asynchronously. The conventional way of modeling speech as non-overlapping segments (e.g. in units of phonemes) may be very inaccurate around the transition regions of speech units. Thus, modeling speech as several parallel continuous articulatory features with asynchronous transition might be more suitable [4]. It is also suggested that the AFs are more robust towards cross-speaker variation and signal distortions such as additive noise, which better reflect real-world conditions.

Practically, AFs have desirable properties [3] to be incorporated in the automatic speech recognition systems (ASR). The AF set is sufficiently small, compared to the commonly used basic speech units such as phonemes. This small size enables maximal sharing of acoustic data by different features, and hence requires minimal amount of training data. On the other hand, the AF set is sufficiently large that combinations of articulatory features provide consistent and deterministic mapping to speech units such as phonemes.

Despite the advantages of incorporating AF into ASRs, there is no general speech corpus with good AF transcription in existence. Experiments with AF set usually have to be carried out by a heuristic mapping from phonetic transcriptions.

[*] This is a Final Project Report for EE225D, Prof. N. Morgan and Prof. B. Gold, EECS, UC. Berkeley, Spring 1999.

[†] Thanks to Dr. L. Shastri, K. Kirchoff, and members of ICSI Realization groups for comments, suggestions, and assistance.

In the past, researchers have attempted various models for AF extraction and modeling. Richards et. al. [12] have used articulatory codebook for point-to-point mappings from the articulatory to acoustic domains. Deng and Sun [4][3]used HMM for overlapping AF representation. Kirchoff [10] developed MLPs for AF extraction. King et. al. [8] developed an articulatory featured syllabic model of speech recognition system.

# 2 Temporal Flow Model

In this work, we developed an AF extraction scheme based on the Temporal Flow Model (TFM) of Watrous and Shastri [15]. TFM supports arbitrary link connectivity across multiple layers of nodes, admits feedforward as well as recurrent links, and allows variable propagation delays to be associated with links (cf. Figures 2 and 3). The recurrent links in TFM provide a means for smoothing and differentiating signals, measuring the duration of features, and detecting their onset. The use of multiple links with variable delays allows the system to maintain context over a window of time and thereby carry out spatio-temporal feature detection and shift-invariant pattern matching. In combination, the use of recurrent links and variable propagation delays provide a rich mechanism for simulating such properties as short-term memory, integration and context sensitivity — properties that are essential for processing time-varying signals. In the past TFM has been successfully applied to a number of tasks including phoneme discrimination [16][17], syllabic segmentation [14], and hand-printed digit recognition [13].

# 3 Experiments

## 3.1 Numbers95 Corpus

The experimental test-bed was a subset of the Numbers95 corpus [2] containing "fluent" numbers such as are spoken in the context of household addresses, over telephone bandwidth. Each utterance in this corpus was labeled and segmented at the phoneme level. A total of 33 different syllables and 34 different phonemes occur in this corpus. Despite the restricted size of the lexicon, the corpus contains speech spoken by a large number of individuals (of both genders) spanning a wide range of geographical dialects, speaking rates and variable utterance lengths. Our training set consists of 600 utterances, 100 of which were used for cross-validation. The test set consists of 500 utterances distinct from those in the training set.

## 3.2 Front-end Processing

The speech waveforms in the corpus were first processed into a modulation-filtered spectrogram (MSG) representation [5][9]. This representation encodes the speech signal in terms of low-frequency energy ($<$ 16 Hz) across time and frequency. It was shown that significant alteration of the modulation spectrum has a deleterious effect on speech intelligibility [6]. For the current study the spectrum was partitioned into 13 discrete, critical-band like channels, over which the MSG was computed using a 250-ms, Hamming window with a slide interval of 10-ms. Other front-end processing schemes such as RASTA [7], may also be used, but were not included in this study due to time and resource constraints.

## 3.3 Articulatory Feature Set

There were many choices of AF sets to be used. For this work, we adopted a particular AF system with five categories of orthogonal articulatory dimensions. The AF categories and features are shown in Table 1. The AF labels for the training and test utterances were obtained by a heuristically defined deterministic mapping from phoneme to feature in the category.

## 3.4 Overall Experiment Setup

The two phases of the overall experiment setup is shown in Figure 1. In the first phase, one TFM neural network was constructed for each feature category. The input to each of the TFM was a vector of 13 MSG values for each

| Feature Category | Size | Features |
|---|---|---|
| Voicing | 3 | +voice, -voice, silence |
| Manner | 7 | stop, vowel,fricative, approximant, nasal, lateral, silence |
| Place | 10 | dental, labial, coronal, palatal, velar, glottal, high, mid, low, silence |
| Front-Back | 4 | front, back, nil, silence |
| Lip Rounding | 4 | +round, -round, nil, silence |

Table 1: Feature categories, number of features in each category, and features.

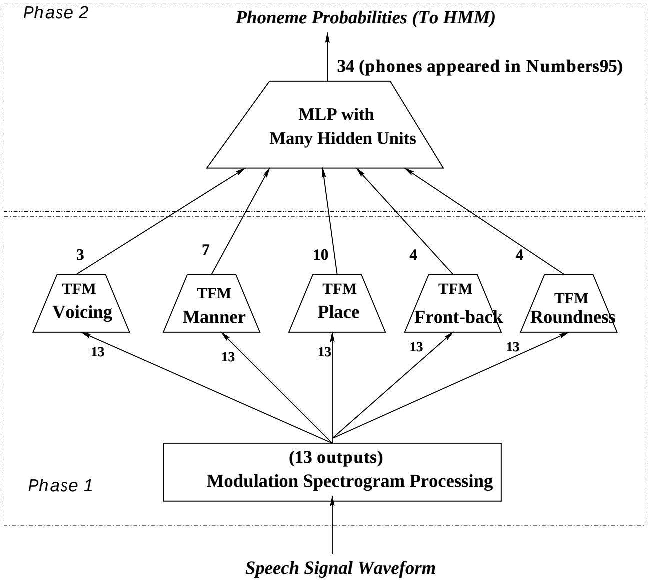

Figure 1: The overall experiment setup shown in two stages. See text for detail.

time frame. Each TFM had one output node for each of the possible features in the corresponding category, and the target value for each output node was either one (for an on-feature) or zero (for an off-feature). Each of the TFMs was trained and tested seperately, and essentially performing an one-out-of-N classification. Since we have used either sum of squared error (SSE) or cross-entropy criteria, upon convergence[1], the network outputs approximated the posterior probabilities of each feature given MSG inputs. Taking the feature with highest output value therefore gave the optimal decision under Bayesian error framework [11].

In the second phase, the 28 outputs of all TFMs were combined and fed into a single-hidden-layer multilayer perceptrons (MLP) with 400 hidden units and nine frames of context window, to esitmate the posterior probabilities of phoneme for each time frame. Again, the phoneme that corresponds to the output node with highest value was picked as the label of the frame for determining phoneme frame accuracy. The vector of all MLP outputs can be directly plugged into the HMM training in a full HMM/ANN-based ASR.

### 3.5 Network Architecture

Two distinct TFM network configurations were investigated, one with global connectivity (Figure 2), the other with tonotopic connectivity (Figure 3). Both configurations contained an input layer, two hidden layers (H1 and H2), and an output layer. The input layer in both configurations contained 13 nodes - one for each of the eleven MSG features. The two network configurations differed, however, in (a) how the input layer was connected to H1 and (b) the density of lateral connections within H1. In the global configuration, all

[1] Here we assume a global minimum or a "fairly close" local minimum was obtaied.

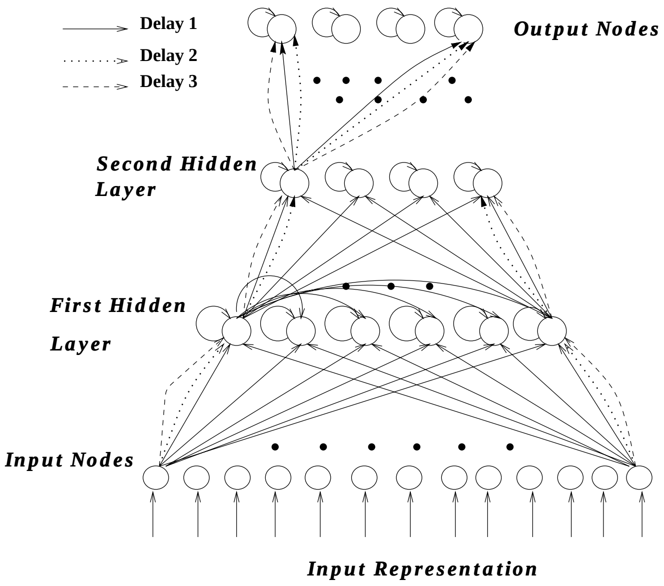

Figure 2: TFM Network Structure: a typical global model with 13 inputs and 4 outputs. Heavy dots in the figure denotes omitted links.

input nodes were connected to all H1 nodes and all H1 nodes were densely connected via lateral links. In the tonotopic configuration, H1 nodes were divided into distinct groups, each receiving activation from a small number of adjacent input nodes (i.e., channels). Nodes within a group were densely connected, but nodes across groups had only sparse interconnections. In both configurations, H1 nodes were fully connected to H2 nodes which, in turn, projected to the output node.

### 3.5.1 Global Model

Figure 2 shows a typical configuration of the global model used for the “Lip-rounding” feature category. The model has 13 input nodes, each receiving an MSG feature. H1 and H2 consist of hidden nodes with self-recurrent links. Between each input node and each node in H1 there are three separate links, each with a different propagation delay (1, 2, 3). Nodes within H1 are also connected with lateral links. Between each node in H1 and each node in H2, there are three links with delays 1,2, and 3, respectively. Nodes in H2 are connected to the output nodes via a similar constellation of links. In general, the number of links, propagation delays and the number of hidden nodes can vary depending on the task.

### 3.5.2 Tonotopic Model

Figure 3 shows a typical configuration of the tonotopic model used for the “Lip-rounding” feature category. The hidden nodes in H1 are divided into four distinct groups. Each of these receives activation from four adjacent input nodes. The input nodes of adjacent groups overlap by a factor of one (i.e., the “receptive fields” of two adjacent groups overlap by 1). An H1 node receives three links with propagation delays of 1, 2 and 3, respectively, from each input node in its receptive field. All nodes within a group are fully connected with links of different propagation delays. Nodes across groups are also connected via links of different propagation delays, but these links are quite sparse. The H2 nodes receive three links from each H1 node with propagation delays of 1,2,and 3, respectively. H2 nodes are also fully linked to the output nodes in a similar manner. In general, the size of, and the overlap between, the receptive fields of H1 nodes, the number of nodes within each group in H1, the number of links, propagation delays and the number of H2 nodes can vary depending on the task.

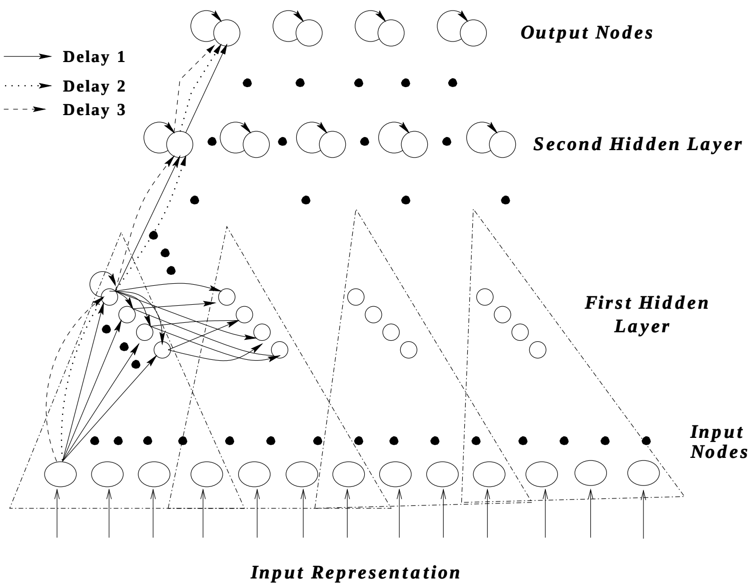

Figure 3: TFM Network Structure: a typical tonotopic model with 13 inputs and 4 outputs.Heavy dots in the figure denotes omitted links.

## 3.6 Training and Testing Procedures

The training of the TFM networks were carried out with Gradsim [18] a gradient optimization package that supports both time-delayed and recurrent links, and almost any arbitrary link connectivities. For all experiments, we have used the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method, a second order gradient-based weight optimization scheme. For the error criteria, we have experimented with both sum of squared errors (SSE) and cross-entropy. In almost all cases, the cross-entropy criteria appeared to be superior, in both convergence time and network quality. It should be noted that the Gradsim was not optimized for the current task (e.g. it cannot read *pfiles* directly), and the training became a bottle-neck of the entire experiments and consequently limited the size of training set to be used.

### 3.6.1 Two Stage Training for Difficult Categories

During training the TFMs, especially for the categories with large number of features (e.g. "place" has 10 possible features), a few features with large number of examples in training set seemed to dominate (very low false negative and very high false positive responses). Features with very small number of examples in training set tended to have very high false negative responses. This was possiblly due to an insufficient complexity of network parameter space and abundant strong attractors in local minima around error regions where high prior features dominated. To remedy this without largely increasing the network size and training time, we devised a two-stage training procedure for feature categories with large number of features. For example, in the case of "place", we first heuristically created coarser distinctions by grouping several features together (e.g. the features "labial" and "dental" were grouped together). This resulted in a reduction from ten to five possible features for the "place". We trained a TFM with five output nodes for the "place" feature category. In the second stage, we created a TFM with ten output nodes corresponding to the original ten possible features of "place", and 18 inputs, 13 for MSG vector and 5 for the outputs from the network trained in the first stage. This resulted in a better performance as shown in Table 2. The trade-off for the gained performance was a slightly longer training time and larger number of parameters (two TFMs instead of one).

| Training | True Positive Rate for each feature | | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | dental | labial | coronal | palatal | velar | glottal | high | mid | low | silence | Frame Acc. |
| 1-stage | 14.46 | 38.96 | 73.91 | 59.45 | 53.46 | 0.00 | 74.14 | 64.48 | 73.49 | 83.99 | 69.32 |
| 2-stage | 4.14 | 52.57 | 73.84 | 61.22 | 66.51 | 0.00 | 77.56 | 67.62 | 78.82 | 84.41 | 72.46 |

Table 2: Comparison between performances of one-stage and two-stage training of TFM for "place" category. All numbers are in percentage.

| Feature | TFM Type | TFM Accuracy | MLP 50 HU | MLP 100 HU | MLP 200 HU | MLP 400 HU |
|---|---|---|---|---|---|---|
| Voice | tonotopic | 88.60 | 84.83 | 84.98 | 85.06 | 85.11 |
| *#params* | 1-stage | *(1186)* | *(6100)* | *(12200)* | *(24200)* | *(50000)* |
| Manner | global | 78.93 | 75.61 | 77.26 | 76.84 | 76.88 |
| #params | 2-stage | *(2746)* | *(6100)* | *(12200)* | *(24200)* | *(50000)* |
| Place | global | 72.46 | 68.11 | 69.00 | 70.23 | 70.69 |
| #params | 2-stage | *(4410)* | *(6100)* | *(12200)* | *(24200)* | *(50000)* |
| Fr-back | global | 79.42 | 75.93 | 77.19 | 77.37 | 75.84 |
| #params | 1-stage | *(1168)* | *(6100)* | *(12200)* | *(24200)* | *(50000)* |
| Round | global | 80.05 | 75.48 | 77.35 | 77.77 | 78.42 |
| #params | 1-stage | *(1168)* | *(6100)* | *(12200)* | *(24200)* | *(50000)* |

Table 3: Frame accuracy of features for the best TFM networks and MLP networks. Accuracies are given in percentage.

## 4 Results and Performance Evaluation

Table 3 shows the frame accuracy for each feature category using the TFM networks, and the number of parameters in the models. As a comparison, we also created single-layer multilayer perceptrons (MLP) with 50 to 400 hidden units (HU), to perform the same AF extraction tasks. The MLPs were trained using QuickNet, and a nine-frame context window was used in all experiments. The results showed that the TFMs gave comparable or better results for all feature categories with much smaller number of parameters as the MLP counterparts. In these experiments, the performances of global and tonotopic networks were not observed to differ significantly.

We computed the frame accuracy of phonemes at the output of the MLP that took as inputs the posterior probabilities for all features estimated by the TFMs. The frame accuracy of phonemes on the test set was 71.16%. As a comparison, we also trained an MLP with 400 hidden units to estimate the posterior phoneme probabilities directly from MSG vectors without going through the AF extraction. The frame accuracy of phonemes for the MLP-only model was 69.02%.

Although the frame accuracy of phonemes in the models with and without AF extraction appeared to be close, their confusion matrices looked considerablly different. This prompted for a combination of the two results for a possiblly better performance. Two simple combination rules were tried. The first was taking the sum of posterior phoneme probabilities from the two models, and the second was taking the product of these probabilities. Indeed, significant improvements were obtained, where the sum rule yielded a frame accracy of 74.45%, and the product rule 74.86%.

## 5 Network Analysis

One advantage of the compact size of the TFMs is that it may be easier to perform post-training analysis on the network, to understand how and what the network have actually learned. The motivation for doing such analysis can be finding hints for network pruning and discovering new sub-features important for the feature

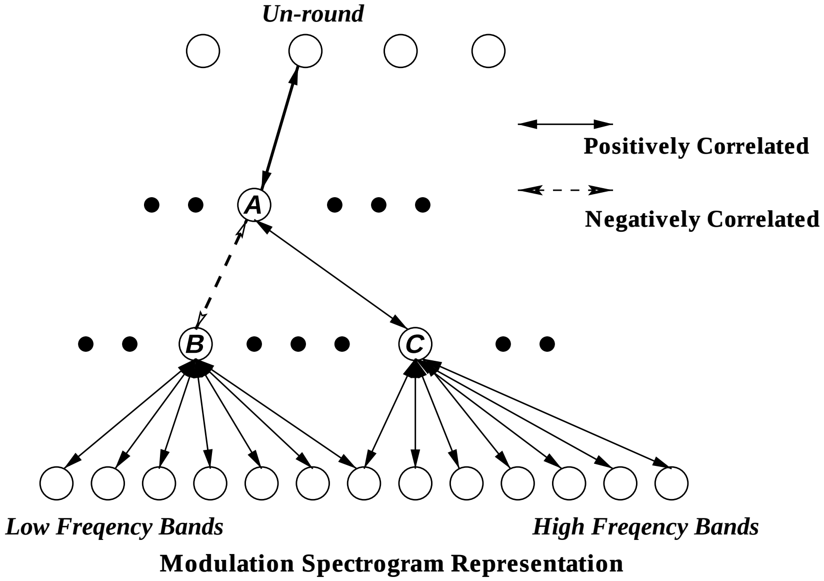

Figure 4: Network node activation correlation pattern for “un-round” in Lip-Rounding feature category.

extraction task. In this study, we have used a simple technique of computing correlation of node activations as a trained network responded to an utterance. We present examples of findings from this analysis below.

## 5.1 Learning in the Hidden Layers

In the TFM for the “lip-rounding” feature, we have discovered some interesting correlations between the output nodes, hidden nodes, and input nodes. As shown in Figure 4, the output node “un-round” was positively correlated with some node (A) in the second hidden layer; node A was found to negatively correlate with some node B in the first hidden layer, and positively correlate with some node C in the first hidden layer. Further more, node B appeared to be positively correlated with input nodes in the lower freqency bands, and un-correlated with inputs nodes in the higher frequency bands. On the other hand, node C appeared to be positively correlated with input nodes in the higher freqency bands, and un-correlated with inputs nodes in the lower frequency bands. Such a correlation pattern suggested that node A was essentially computing some enery difference between the higher and lower frequency bands in the input signal. Why was this correlation pattern an useful feature for distinguishing “round” and “un-round”?

Figure 5. shows the common positions of first and second formants of some cardinal vowels (CV) [1]. The CVs on the left of horizontal axis are more “unround”, and the ones on the right are more “round”. From this figure, one can easily see that a more positive difference between the energy in higher and lower freqeuncies coresponds to a more likely “un-round” CV. Of course, this was probablly not the only cue that the network used to distinguish “un-round” from “round”. Further analysis may reveal other interesting correlation patterns.

Another example of the node activation analysis was on the TFM trained for the “manner” feature category. It was found that a node in the second hidden layer which was highly correlated with the output node for “approximants”, seemed to correlate with an upward shift of spectral energy from lower to higher frequency regions in successive time frames. Whether this was a potential cue for finding “approximants” is subject to further investigation.

## 5.2 Context Window Analysis

As previously mentioned, the multiple time-delayed links and recurrent links in the TFM provide a flexible context window. To better understand the relationship between the network architecture and context window

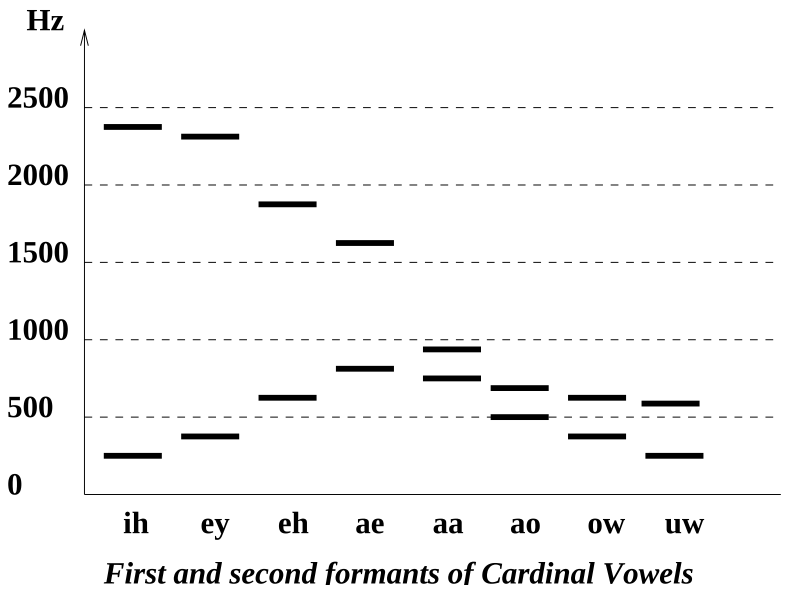

Figure 5: First and second formants of some Cardinal Vowels.

size, analysis was performed on trained TFM network by providing a single impulse in a single input channel at some time frame. Figure 6 shows node activations when all recurrent links were removed from the network (multiple time-delayed links were intact), and Figure 7 shows same node activations with all recurrent links reinstated. For the network with recurrent links absent, the context window size can be calculated as follows:

Let $min$ be the minimum delay for signal to reach output nodes from input nodes. Let $max$ be the maximum delay for signal to reach output nodes from input nodes. The context window size is then $max - min + 1$. In the network for producing Figure 6, $min = 3$ and $max = 9$, and the context window size is therefore 7. This can be verified by looking at the output node activations (between 3 steps and 9 steps after the input activation).

Clearly, in the network with recurrent links, the context window size can be arbitrarily large, and currently it remains difficult to fully characterize the stability and capability of the recurrent system.

## 6 Future Work

A number of extensions to the current experiments can be performed. The phonetic transcription of the Numbers95 corpus contains many inaccuracies. For a better evaluation of the models, we should perform the experiments on a corpus with more accurate transcription. One such corpus available is the TIMIT.

The current study was limited to the extraction of AFs and finding the posterior phoneme probabilities of each frame from the AFs. It would be interesting to embed the AF extraction system into a full HMM/ANN-based ASR system for an evalution of the system at a word recognition level. Also the performance of the AF extraction system may improve as a result of iteratively embedded training.

The current experiments only used MSG as front-end processing. It is reasonable and interesting to try out other processing methods, such as RASTA-plp, MFCC, etc. Better performance may also be expected for combined results of different front-end processings.

One potential advantage of AF set is the expected robustness and invariance across speaker and acoustic conditions. Kirchoff [10] has demonstrated that incorporating AF set in the ASR was particularly benefitial to reverberant and noisy speech. It would be interesting to see the results of the TFMs under degraded signal conditions.

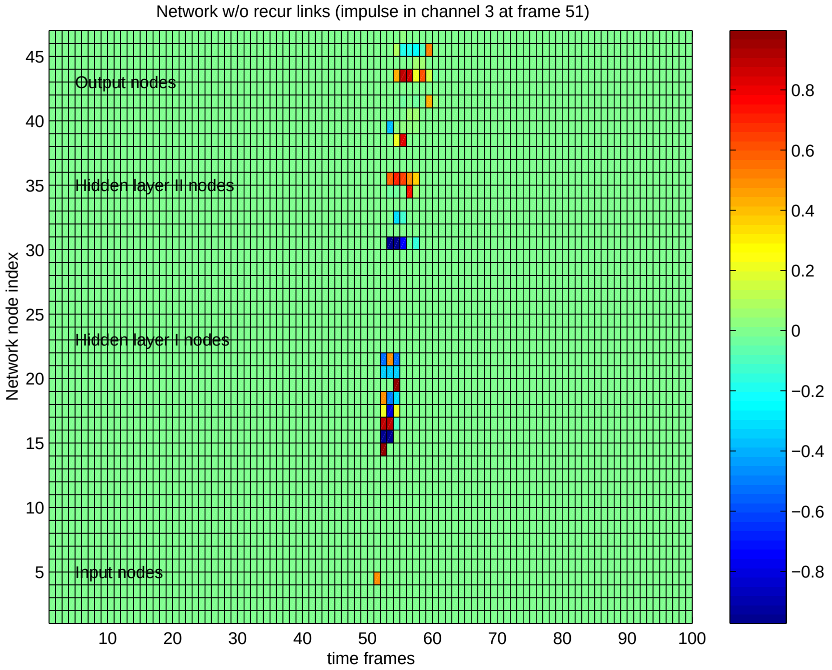

Figure 6: Node activations to an impulse in channel 3 at frame 51, for a TFM with all recurrent links removed.

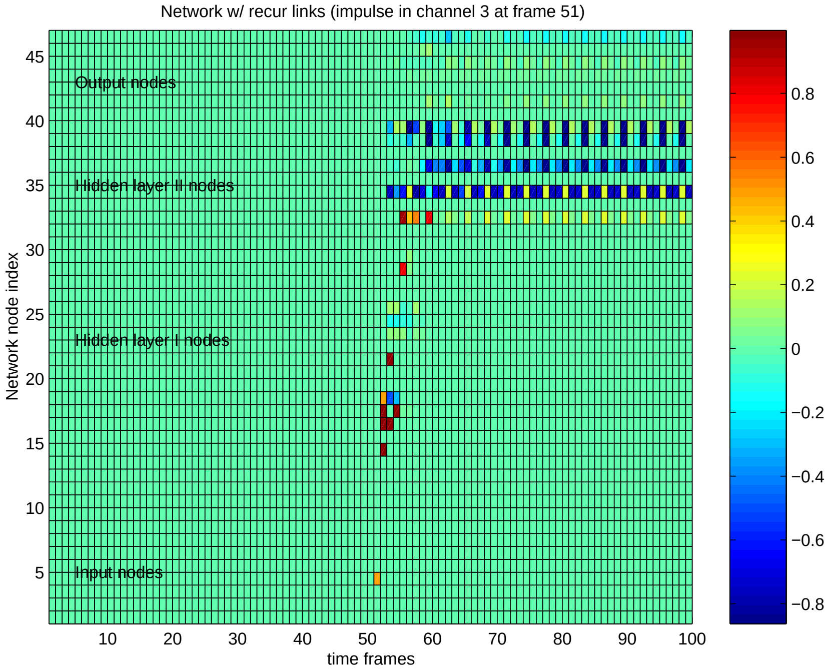

Figure 7: Node activations to an impulse in channel 3 at frame 51, for a TFM with all recurrent links intact.

# 7 Conclusion

This paper presented the experiment of extracting AFs from pre-processed speech signals using TFM neural networks. Results show that the TFMs were able to obtain good performances, comparable or better than that of MLP counterparts with much larger number of free parameters. We also applied the outputs from the TFMs in the estimation of posterior phoneme probabilities with large MLP, and achieved reasonable results. Combination of the results from the TFM/MLP system and MLP-only system gave significant improvements in the phoneme frame accuracy. We also analyzed the trained TFM network and discovered interesting node activation correlation patterns, corresponding to useful cues for AF extraction. The effect of recurrent links and mutiple time-delayed links of TFMs on the context window size was also investigated.

# References

[1] *A practial introduction to phonetics*, Oxford University Press, New York, 1988.

[2] Center for Spoken Language Understanding, Dept. of Compu ter Science and Engineering, Oregon Graduate Institute. Numbers corpus, R elease 1.0. 1995

[3] Deng, L. and Sun, D., Phonetic classification and recognition using HMM representation of overlapping articulatory features for all classes of English sounds. *Proceedings of ICASSP 1994*, 45-48, Adelaide, Australia, 1994.

[4] Deng, L. and Sun, D., A statistical approach to ASR using atomic units constructed from overlapping articulatory features. *JASA*, 95:2702-2719,1994.

[5] Greenberg, S. and Kingsbury, B. The modulation spectrogram: In pursuit of an invariant representation of speech, *ICASSP-97, IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1647- 1650. 1997.

[6] Greenberg, S. Speaking in shorthand - A syllabl e-centric perspective for understanding pronunciation variation, *Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Kekrade (Netherlands), pp. 47-56. 1998

[7] Hermansky, H. and Morgan, N., RASTA processing of speech. *IEEE Transaction on Speech and Audio Processing*,2:578-589, 1994.

[8] King, S.,Stephenson, T., Isard, S.,Taylor,P., and Strachan, A., Speech Recognition via phonetically featured syllables. *Proc. ICSLP98*, 1998.

[9] Kingsbury, B., Morgan, N. and Greenberg, S. Rob ust speech recognition using the modulation spectrogram, *Speech Communication*, 25, 117-132. 1998.

[10] Kirchoff, K., Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments. *Proceedings of ICSLP 1998*, 1998.

[11] Richard, M., and Lippmann, R., Neural network calssifiers estimate Bayesian a posteriori probabilities, *Neural Computation*, 3:461-483, 1991.

[12] Richards, H., Mason, J., Hunt,M., and J. Bridle, Deriving articulatory representations of speech. *Proc. ICSLP96*, 2:1233-6,1996.

[13] Shastri, L. and Fontaine, T. Recognizing handwr itten digit strings using modular spatio-temporal connectionist networks, *Connection Science*, 7(3,4), 211-245. 1995.

[14] Shastri,L., Chang, S., and Greenberg, S., Syllable Detection and Segmentation using Temporal Flow Neural Networks, *to appear in the Proceedings of ICPhS99*, San Francisco, August, 1999.

[15] Watrous, R. L. and Shastri, L. Learning phonetic features using connectionist networks: An experiment in speech recognition, Tech. Report, MS-CIS-86-78, University of Pennsylvania.1986.

[16] Watrous, R. L. and Shastri, L. Learning phonetic features using connectionist networks. *Proceedings of IJCAI-87, the Tenth International Joint Conference on Artificial Intelligence*, Milan, Italy, August 1987. pp. 851-854.

[17] Watrous, R. L. Phoneme discrimination using connectionist networks, *Journal of Acoustic Society of America*, 87, 1753-1772.1990.

[18] Watrous, R. L. GRADSIM: a connectionist network simulator using gradient optimization techniques, Report, Siemens Corporate Research, Inc., Princeton, New Jersey. 1993.