Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Nonlinear Dynamic Neural Network for Text-Independent Speaker Identification using Information Theoretic Learning Technology

Bing Lu, Walter M. Yamada, and Theodore W. Berger

Abstract— **In this paper we present a novel design for a nonlinear dynamic neural network to implement text-independent speaker recognition without the benefit of exact voice signatures. The dynamic properties between the input neuron and the output neuron make use of a nonlinear high-order synaptic neural model with memory of previous input signals. The dynamic neural network is realized in the short-term-frequency long-term-temporal domain. Informatics metric is used to overcome the challenge of performing blind learning for the nonlinear network. The goal of this study is not only to improve the recognition performance but also to amplify the distinctiveness among different speakers.**

I. Introduction

NORMAL speech not only conveys information via words, but also contains information such as the gender and the identity of speakers. [1] and [2] provide a good overview of technologies, applications, and challenges in speaker recognition. Many studies have attempted to use hidden Markov models (HMM) as a decoding algorithm for speech recognition and speaker identification. However, this technique assumes independence of consecutive feature vectors, and is a short-windowed analytical technique, and when compared to natural listener performance has inferior classification ability. The basic challenge of speaker recognition is that the exact desired 'voice signature' of the incoming stimuli is inherently unknown to the speaker recognizer. However, perhaps utterances recorded over a long time would provide more attributes of voice signature than can be captured in a short-time duration. In fact, [1], [2] raise the possibility that higher level cues such as prosodic features or other long-term signal measures may improve the accuracy of speaker recognition.

As pointed out in [3] and [4], mammalian brains process auditory signals in a two-dimensional way (both time and frequency), and the receptive temporal field is extended up to the order of hundreds of milliseconds. In [5], a time-frequency deformation is proposed which employs scalar weights as the connections between 'neurons'. However this and most models to date employ static architectures that do not address the dynamic behavior of neurons from the perspective of biophysiological reality.

In the present paper we address the issues of incorporating higher level cues as well as biophysiological realism while implementing the general computational purpose of speaker recognition. The rest of this paper is organized as follows. In section II, a simple review of high-order dynamic synapse (HODS) model is described. In section III, we build a short-term-frequency (STF) long-term-temporal (LTT) two-dimensional neural network by applying dynamic neurons as computation units. The nonlinear temporal transformation property is illustrated in subsection IV-A. The importance of information theoretic metric on an unsupervised learning scheme is described in subsection IV-B. Simulation results are given in section V, where it is shown that the designed model can offer a high classification rate along with increased distinctiveness between speakers. Some highlighted conclusions are discussed in section VI.

Authors are in Department of Biomedical Engineering, University of Southern California, Los Angeles, CA 90089, USA (contact email: blu@usc.edu)

II. High-order dynamic synapse model

In [6] it is shown that considering the biophysical mechanisms of synapses such as the connections between neurons in a biological neural network and the properties of cell membranes results in an alternative approach to understanding neural computation that warrants attention. In [7], a high-order dynamic synapse model is developed which describes a queue model composed of four serially connected neurotransmitter pools at the pre-synapse. Static or time-varying transfer rates describe the stationary or non-stationary neurotransmitter vesicle transmitting events among four neurotransmitter pools.

In [7] it is explained how neurotransmitter release events are affected by facilitation and depression in a third-order fashion between the pre-synapse and the post-synapse. This model has been tested upon the experimental data recorded at hippocampal Schaffer Collateral – CA1 cells. Most notably, the HODS model realizes timing relations between the current input spike and previous input spikes. With a set of fixed parameters, one high-order dynamic synapse can be considered to be a dynamic filter capable of yielding output responses sensitive to various inputs and their particular history. Thereby the current post-synaptic response $y(t_n)$ is calculated in terms of previous pre-synaptic inputs $\{x(t)|t \leq t_n\}$ as

$$y(t_n) = f(x(t), \mathbf{w}|t \leq t_n), \tag{1}$$

where $f(\cdot)$ denotes the HODS nonlinear transformation function, and the vector $\mathbf{w}$ represents all seven parameters of one HODS model (for details, please refer to [7]). We will impose this temporal property on our nonlinear dynamic neural network design.

III. Nonlinear Dynamic Neural Network

Physiological studies of the mammalian auditory cortex in [3], [4] have determined that neurons in the brain process both time and frequency components of signals. This

has motivated us to consider signal processing in a two-dimensional way, instead of single-dimension analysis (either time or frequency). Moreover, as declared in [3], [4], the receptive field for auditory signals can be extended up to the order of hundreds of milliseconds. Keeping this in mind, we designed a scheme to retrieve high-level, time-independent acoustic information that analyzes the signal over 'all' time, i.e., we designed a functional transformation to convert the temporal input into a multi-dimensional output function in time.

We choose to use a short-term-frequency long-term-temporal neural network, and as such, output of the network is a function of the input data over a period of time. Each acoustic signal is first divided into a series of non-overlapping frames, each having a duration of T_f. T_f is normally chosen on the order of tens of milliseconds. Due to the large variability of acoustic signals, it is fairly typical in acoustic signal processing to perform some form of feature extraction to reduce the signal variability and to decrease the degrees of freedom using a method such as Mel frequency cepstral coefficient (MFCC) computation. MFCC is similar to the cochlea of the human ear in that it performs a quasi-frequency analysis. MFCC acts on each separate frame in a method known as short-term spectrum calculation.

Based on MFCC features, a nonlinear dynamic neural network is employed to explore the timing course at each quefrency channel (in general, inverted log-frequency transform is defined as quefrency). The suggested design is quite different from typical artificial neural networks as the design addresses neural dynamics. Despite the exact details of biological learning procedures being unknown, dynamic connections are stated to be existent between neurons in a biological neural network. By applying nonlinear dynamic neural functions, we are taking a step towards investigating the processes of biological learning.

A two-dimension network structure is built using high-order dynamic synapses as computation units. The purpose of the structure is to predict long-term spectrum and to extract further temporal features over consecutive frames. The pre-synapses and their corresponding post-synapses are aligned in the quefrency coordinate to represent the Q-by-1 feature vector associated with short-term spectrum computation. One pre- and post- synaptic pair is designated to denote the time coordinate. We define the temporal memory being M. From the nonlinear function eq. (1), the output $y_q(n)$ at quefrency q is

$$y_q(n) = f(\mathbf{x}_q(n), \mathbf{w}_q), \quad q \in [1, Q] \tag{2}$$

where $\mathbf{x}_q(n) := [x_q(n), x_q(n-1), \ldots, x_q(n-M+1)]^T$, and $[\cdot]^T$ represents vector transpose. Thus the output effectively embodies temporal traits by integrating previous temporal features. An overview data flow is shown in Figure 1. Q short-term input vectors are first interleaved. At each quefrency, one nonlinear dynamic neural model accounts for timing variation over M frames. Having experienced the dynamic network, the output data is interleaved back so that it can be used to yield further spectrum demonstration.

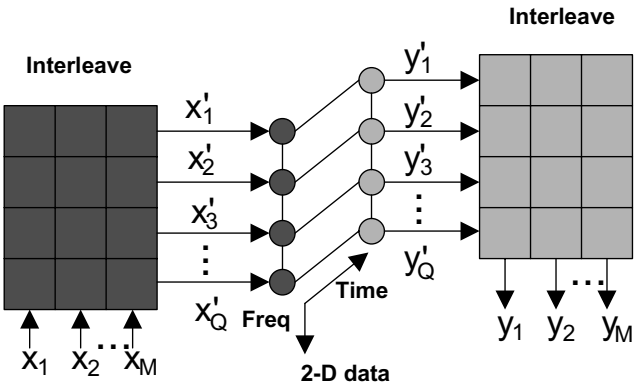


Fig. 1. An overview of data flow process

This design is essentially different from general artificial neural networks, which generate a weighted summation output from activated responses of previous layers. Instead, our proposed neural network yields the nonlinear response from previous layers individually at each path (quefrency). This separability allows us to train the proposed dynamic network in parallel as described in subsection IV-B. Furthermore, relying on the HODS model in [7], we find that the output response is actually a quasi-exponential function with memory of previous events. This nonlinear function consists of several exponential or quasi-exponential functions to represent dynamics, instead of the scalar weights generally employed in time delay neural networks.

IV. Text-Independent Speaker Identifier Based on Information Theoretic Metric

A. Nonlinear temporal transformation

A hallmark of human learning is the ability to classify acoustic signals. We can distinguish utterances by activating temporal dynamic neural structures according to prosodic cues of specific speakers. The biophysical mechanisms underlying this ability are unknown, but generally, classification ability can be described as a filtering process: signals associated with one speaker are 'amplified' whereas signals associated with another speaker are 'attenuated'. As a result, signals from different speakers can be separated farther away such as illustrated in Figure 2, which displays nonlinear transformation from x feature space to another y feature space.

The goal of our design is to establish a selective association such that dynamic neural models can adaptively approach the feature space of one specific speaker. To highlight the decoding selectivity, a conceptual example of a decoding filter is shown in Figure 3. Two different input signals x_1 and x_2 with distance 5.6 are provided on the left. By training the filter f_1 to ideally benefit the selectivity of the first signal, the filtered responses for two signals in y space are illustrated on the right with distance 59.7. As a result of decoding selectivity, the output response y_1 is comparatively amplified while the output response y_2 is comparatively attenuated with the increased distance. So the filtered outputs are easier

to discriminate than the original signals. This is a simple example of how the decoding filter acts on the original data. Speaker identification is a more sophisticated case.

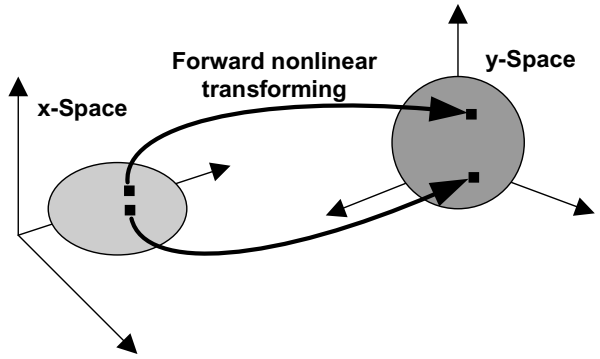


Fig. 2. Nonlinear transformation illustration

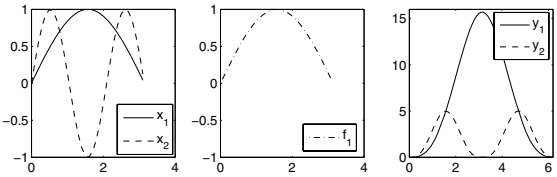


Fig. 3. An example of a decoding filter in continuous time

B. Information theoretic metric for unsupervised learning

In order to discriminate speaker attributes, we aim to investigate high-level acoustic features, i.e., we focus more on how the person speaks and focus less on what is exactly said. Utterances indicate a speaker's identity information at various levels. On the other hand, the 'exact' voice signatures of different speakers are generally not predefined. Our proposed speaker detector possesses two kinds of blind processes. First, semi-supervised learning exists between speakers as no exact voice transcripts are known for each speaker though the signals are labelled to speakers. Second, unsupervised learning of acoustic signals exists inside each speaker's training, i.e., to design a text-independent speaker recognizer, we propose to separate high-level acoustic signals (speaker identity information) from low-level acoustic signals (texts). We will draw upon information theoretic metric to seek the resolution of above blindness and to train the dynamic temporal neural network to selectively corroborate the discriminability between speakers.

Specifically, supposing the 'real' voice signature $s(n)$ is reflected over M frames, we assume high-level acoustic signals coherent and the divergent component $v(n)$ noncoherent. We also take the assumption that high-level acoustic information is independent of and additive with low-level information at quefrencies. Ignoring q for simplicity, the nonlinear function (2) is approximately expressed as

$$\begin{aligned} y(n) &= f(\mathbf{s}(n) + \mathbf{v}(n), \mathbf{w}) \\ &= f'(\mathbf{s}(n), \mathbf{w}) + f'(\mathbf{v}(n), (\mathbf{s}(n), \mathbf{v}(n)), \mathbf{w}), \end{aligned}$$

where $\mathbf{s}(n), \mathbf{v}(n)$ are vectors over memory M, $f'(\mathbf{s}(n), \mathbf{w})$ denotes all orders of nonlinear responses from stimuli $\mathbf{s}(n)$, and $f'(\mathbf{v}(n), (\mathbf{s}(n), \mathbf{v}(n)), \mathbf{w})$ represents all orders of nonlinear responses from stimuli $\mathbf{v}(n))$ and from interact part $(\mathbf{s}(n), \mathbf{v}(n))$.

To approach the purpose of extracting speaker's common features out of acoustic signals, we turn toward augmenting the item $f'(\mathbf{s}(n), \mathbf{w})$ associated with inhibiting the item $f'(\mathbf{v}(n), (\mathbf{s}(n), \mathbf{v}(n)), \mathbf{w})$. Shannon's distortion rate theory [8][p. 337] states that the decoding and classification performance is optimized if the capacity $\log(1 + f'^2(\mathbf{s}(n), \mathbf{w}) / f'^2(\mathbf{v}(n), (\mathbf{s}(n), \mathbf{v}(n)), \mathbf{w}))$ is maximized, where $\log(\cdot)$ is a monotonic function that does not modify the metric variation trend. With the assumptions in the last paragraph, we suppose $f'^2(\mathbf{s}(n), \mathbf{w})$ is coherent and ergodic, and converges to the mean value of the output response; we assume $f'^2(\mathbf{v}(n), (\mathbf{s}(n), \mathbf{v}(n)), \mathbf{w})$ is noncoherent and zero-mean. Therefore we derive the following from the capacity definition as the cost metric for training

$$\Upsilon(y) = \log \frac{(\mathrm{E}[y(n)])^2}{\sigma_y^2}, \tag{3}$$

where $\mathrm{E}[\cdot]$ denotes the statistical mean, and σ^2 the variance.

From the perspective that speaker's identity information is contained in each acoustic signal, the temporal dynamic neural network ensembles the high-level signal over a long term, thus factoring out and normalizing various levels of variability. Relying on the cost metric in eq. (3), each speaker's voice signature can be extracted and other low-level acoustic variation is suppressed. As the dynamic network is nonlinear, optimizing parameters is difficult using gradient descent algorithms. We resort to Nelder Mead simplex method to find the optimal/suboptimal solutions.

What's more, when comparing the proposed design with the traditional MFCC high pass filter (HPF) that detects speaker's pitch information and then identifies speakers, we discover three weaknesses for MFCC HPF: i) the threshold between low quefrency and high quefrency is not adaptive; ii) the high quefrency may not be able to completely represent high-level acoustic information; iii) the current output response can be sensitive to its particular temporal history. Hence, a dynamic decoding filter with temporal memory is a better scheme to adaptively adjust pass band and stop band.

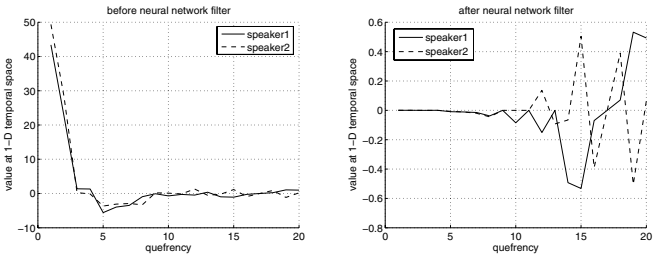
V. Simulation results

In the simulation section, we use TIMIT data and 40 western-dialect speakers to train the designed networks. For each speaker, seven sentences are randomly selected for training and the rest for testing. The speech data is processed by a silence-removing algorithm followed by the application of nonlinear dynamic neural networks. The training procedure is described as maximizing the cost in eq. (3) via Nelder Mead algorithm searching parameter vector $\mathbf{w}$. $T_f = 10$ ms, $M = 30$, $Q = 20$. After about 110 epoches, the training procedure converges. Upon converging, the nonlinear decoding filter expressed by $\mathbf{w}$, the output pattern, and the variance

divergent from the pattern are obtained at each quefrency. One strength stemming from the proposed cost metric is that only one speaker's data is used to train this speaker's dynamic neural network. Thus other speakers' data is not needed during training, which dramatically reduces the training time and complexity.

During the testing stage, the resting data from all speakers are mixed to test every trained neural network. Assume the output response is Gaussian distributed (if it is not Gaussian distributed, central limit theory [8][pp. 191-192] proves it is approximately Gaussian over M frames). In [9], it claims that Gaussian distribution belongs to exponential families. As a result, the designed quasi-exponential neural network and the data that the network is modelling analogously inside the same exponential families. Maximal log-likelihood is applied to operate on testing unknown data. The correct identification rate is about 92-97.5%. This is consistent with the fact that each network has learnt features of its own speaker. Hence, each network with respect to its speaker can be analogously viewed as an independent basis of the speaker's feature space. All networks therefore construct a set of independent bases to span the whole data space.

In addition, we discover that maximizing the cost metric in eq. (3) will remove the common correlation between speakers while strengthening the specific voice signature of each speaker. One example is given in Figure 4. Figure 4 (a) shows MFCC patterns of two male speakers within western dialect. For simplicity, only one-dimension (1-D) temporal space is plotted. After employing dynamic network filtering, their temporal outputs are presented in Figure 4 (b). It can be seen that common features between them are suppressed but their distinct features are preserved.



(a) (b)

Fig. 4. (a) MFCC patterns before dynamic network filtering; (b) MFCC patterns after dynamic network filtering

To make the above description more understandable, the following three equations are defined to demonstrate the decoding gains of dynamic networks

$$G_1 = \frac{||\mu_{y^a} - \mu_{y^b}||}{||\mu_{x^a} - \mu_{x^b}||}, \tag{4}$$

where a, b represent any two speakers, μ indicates expectation. The higher this gain value, the farther the two speakers' voice signatures and the less possible that the data can be classified into a wrong speaker. Secondly,

$$G_2 = \log\frac{\Upsilon(y)}{\Upsilon(x)}, \tag{5}$$

TABLE I
COMPARISON RESULTS USING HODS NETWORKS

	Before Network	After Network	Decoding gain
$\|\|\mu^a - \mu^b\|\|$	1.4500	2.5100	$G_1 = 1.7310$
Υ(dB)	-10.8306	3.7293	$G_2 = 14.5653$
$I(a, b)$	0.8810	0.5630	$G_3 = -0.3180$

where $\Upsilon(x)$ is defined similar as in eq. (3). The larger this gain value, the less rate distortion is lost. Thirdly,

$$G_3 = I(y^a, y^b) - I(x^a, x^b), \tag{6}$$

where $I(y^a, y^b)$ denotes the cross mutual information after applying dynamic networks, so $I(x^a, x^b)$ represents the cross value before applying networks. This gain indicates the correlation removal between speakers. The smaller this gain value, the more discriminative the two speakers.

The averaged decoding gains after using the dynamic networks are given in Table 1. They obviously show temporal dynamic neural networks can effectively further distinguish speakers, not only can just identify them.

VI. CONCLUSION

Auditory signal processing in the mammalian cortex provides us with some heuristics to explore the speaker recognition task. We contributed original work to develop a nonlinear dynamic neural network using high-order dynamic synapses in STF LTT feature space. Information theoretic metric is applied for unsupervised learning. The proposed model can selectively amplify signals associated with one speaker and attenuate signals associated with another speaker, resulting in a potential improvement on the recognition performance. Simulation results prove that the proposed model can significantly distinguish differences between speakers even if originally they have a high correlation between each other.

REFERENCES

[1] J. P. Campbell, "Speaker Recognition: A Tutorial," *Proceedings of the IEEE*, vol. 85, pp. 1437–1462, 1997.

[2] D. A. Reynolds, "An Overview of Automatic Speaker Recognition Technology," *ICASSP*, vol. 6, 2002.

[3] S. Shamma, H. Versnel, and N. Kowalski, "Ripple Analysis in Ferret Primary Auditory Cortex: I. Response Characteristics of Single Units to Sinusoidally Rippled Spectra," *Aud. Neurosci.*, vol. 1, 1995.

[4] D. Klein, D. Depireux, J. Simon, and S. Shamma, "Robust Spectro-temporal Reverse Correlation for the Auditory System: Optimizing Stimulus Design," *J. Comput. Neurosci.*, vol. 9, 2000.

[5] M. Reyes-Gomez, N. Jojic, and D. P.W. Ellis, "Towards Single-Channel Unsupervised Source Separation of Speech Mixtures: The Layered Harmonics/Formants Separation/Tracking Model," *Research Workshop on Statistical and Perceptual Audio Processing*, Korea, 2004.

[6] C. Koch and T. Poggio, "Biophysics Computation: Neurons, Synapses and Membrances," *Synaptic Function*, ch. 23, 1987.

[7] B. Lu, W. M. Yamada, and T. W. Berger, "Nonlinear Queuing Model for Dynamic Synapse with Multiple Transmitter Pools," *IJCNN*, 2006.

[8] R. G. Gallager, *Information Theory and Reliable Communication*, MIT, 1968.

[9] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman Divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.