MULTI-DISTRIBUTION DEEP BELIEF NETWORK FOR SPEECH SYNTHESIS

Shiyin Kang, Xiaojun Qian and Helen Meng

Human Computer Communications Laboratory,
Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR, China

ABSTRACT

Deep belief network (DBN) has been shown to be a good generative model in tasks such as hand-written digit image generation. Previous work on DBN in the speech community mainly focuses on using the generatively pre-trained DBN to initialize a discriminative model for better acoustic modeling in speech recognition (SR). To fully utilize its generative nature, we propose to model the speech parameters including spectrum and F0 simultaneously and generate these parameters from DBN for speech synthesis. Compared with the predominant HMM-based approach, objective evaluation shows that the spectrum generated from DBN has less distortion. Subjective results also confirm the advantage of the spectrum from DBN, and the overall quality is comparable to that of context-independent HMM.

Index Terms— Speech synthesis, Deep belief network

1. INTRODUCTION

The past decade has witnessed the success of HMM-based text-to-speech (TTS) synthesis [1]. The core underpinning techniques include: (1) the adoption of multi-space distribution HMMs in estimating the statistical behavior of speech parameters [2]; (2) the parameter generation algorithm which uses dynamic features to smooth the originally piece-wise constant speech parameters drawn from the HMM states [3].

This work is our first attempt in using deep belief network (DBN) to synthesize speech. DBN is a probabilistic generative model which is composed of multiple layers of stochastic hidden variables [4]. Previously, DBN has been shown to learn very good generative and discriminative models on high-dimensional data such as handwritten digits [4],facial pictures [5], human motion [6] and large vocabulary speech recognition [7] . This motivates us to use DBN to model the wide-band spectrogram and F0 contour for speech synthesis.

In this paper, the basic linguistic unit is the Mandarin tonal syllable. We keep the speech vocoding framework as in conventional statistical parametric speech synthesis [8]. DBN is used to model the joint probability of syllables and speech parameters.

The major difference between our approach and the prevalent HMM-based approach is: Instead of using a variable sequence of states to represent the time dynamics of each syllable, we evenly sample a fixed number of frames within the delimited syllable boundary with high resolution. Besides, in contrast to previous application of DBN in acoustic modeling for SR, we model the continuous spectrum, discrete voice/unvoiced decision and the multi-space F0 pattern simultaneously rather than only the continuous spectrum [9].

The rest of the paper is organized as follows: We will introduce DBN in the context of speech synthesis in Section 2. The procedures to synthesize speech using a DBN will be described in Section 3. The experiments and results will be shown in Section 4. Finally we present the conclusions and the future directions.

2. MULTI-DISTRIBUTION DEEP BELIEF NETWORK (MD-DBN)

Speech production theory describes how the linguistic message undergoes a series of neuromuscular and articulatory processes before acoustic realization. To mimic this sequential process, we attempt to model speech production with a directed belief network. However, it is difficult to learn a multi-layered belief network layer by layer, which involves inferring the posterior of hidden units immediately above. The insight of [4] states that this inference can be greatly simplified, i.e., just deriving the posterior from an up-pass, if we assume the factorial prior coming from the upper layers is defined by a restricted Boltzmann machine (RBM) which shares the same connection weights with the current layer. This insight also enables a layer-wise greedy construction of a DBN from bottom-up using RBMs as the building blocks.

To build an MD-DBN for speech synthesis, we use three types of RBMs: (1) mixed Gaussian-Bernoulli RBMs, for spectrum, log-F0 and voiced-unvoiced representation with assumed Gaussian or Bernoulli distributions; (2) mixed Categorical-Bernoulli RBMs, for capturing the correspondence between syllable identities and the binary data derived from the speech representation; and (3) Bernoulli RBMs, which are used to encode binary data.

2.1. Bernoulli RBM (B-RBM)

A B-RBM is an undirected graphical model with one layer of stochastic visible binary units $\boldsymbol{v}$ and one layer of stochastic hidden binary units $\boldsymbol{h}$. There is no interaction between units in the same layer and is thus "restricted". It defines the "energy" of a visible-hidden configuration ($\boldsymbol{v}$, $\boldsymbol{h}$) as follows:

$$E(\boldsymbol{v}, \boldsymbol{h}; \Theta) = -\boldsymbol{h}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{v} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{v} - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{h}, \tag{1}$$

where $\Theta = \{\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b}\}$ is the set of parameters of a RBM and Θ will be omitted for clarity hereafter. w_{ij} is the weight of the symmetric connection between the hidden unit i and the visible unit j, while a_i and b_j are their bias terms. The distribution of the ($\boldsymbol{v}$, $\boldsymbol{h}$) configuration is: $\Pr(\boldsymbol{v}, \boldsymbol{h}) = \exp(-E(\boldsymbol{v}, \boldsymbol{h}))/Z$ (Z is the normalization term), i.e., the higher the energy, the lower the probability.

The nice property of this setting is that the two conditionals $\Pr(h_i = 1|\boldsymbol{v})$ and $\Pr(v_j = 1|\boldsymbol{h})$ can be obtained easily using the fact that h_i and v_j can only be either 0 or 1:

$$\Pr(h_i = 1|\boldsymbol{v}) = \sigma(\sum_j w_{ij} v_j + a_i), \tag{2}$$

$$\Pr(v_j = 1|\boldsymbol{h}) = \sigma(\sum_i w_{ij} h_i + b_j), \tag{3}$$

where $\sigma(x) = (1 + e^{-x})^{-1}$.

To optimize the log-likelihood of $\boldsymbol{v}$ in a first-order approach, we need the gradient of $\log \Pr(\boldsymbol{v})$ with respect to any θ in Θ:

$$\sum_{\boldsymbol{h}} \frac{\partial - E(\boldsymbol{v}, \boldsymbol{h})}{\partial \theta} \Pr(\boldsymbol{h}|\boldsymbol{v}) - \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}} \frac{\partial - E(\boldsymbol{v}, \boldsymbol{h})}{\partial \theta} \Pr(\boldsymbol{h}, \boldsymbol{v}). \tag{4}$$

Given the instantiated observation $\boldsymbol{v}$, the expectation of derivatives in the first term in Eqn. (4) can be easily computed. Unfortunately, the second term in Eqn. (4) involves a summation over all possible $\boldsymbol{v}$ and is intractable. A widely applied method that approximates this summation is the Gibbs sampler which (optionally starts from $\boldsymbol{v}$) proceeds in a Markov chain as follows:

$$\boldsymbol{v}^{(0)} \sim \boldsymbol{v}, \qquad \boldsymbol{h}^{(0)} \sim \Pr(\boldsymbol{h}|\boldsymbol{v}^{(0)}); \tag{5a}$$

$$\boldsymbol{v}^{(1)} \sim \Pr(\boldsymbol{v}|\boldsymbol{h}^{(0)}), \qquad \boldsymbol{h}^{(1)} \sim \Pr(\boldsymbol{h}|\boldsymbol{v}^{(1)}); \tag{5b}$$

$$\dots$$

Contrastive divergence (CD) training [4] makes two further approximations: (1) that the chain starts from the clamped $\boldsymbol{v}$ and is run for only k steps (CD-k); (2) the summation is replaced by a single sample. In particular, CD-1 measures the discrepancy between $\boldsymbol{v}$ and its reconstruction to present a direction for optimization. Starting from a training frame $\boldsymbol{v}^{(0)}$, we only sample $\boldsymbol{h}^{(0)}$ in Eqn. (5a) and use the expectations for Eqn. (2) and Eqn. (3) to replace the random samples $\boldsymbol{v}^{(1)}$ and $\boldsymbol{h}^{(1)}$ in Eqn. (5b) for stability.

2.2. Mixed Gaussian-Bernoulli RBM (GB-RBM)

Speech parameters for synthesis include the spectrum and the log-F0 with assumed Gaussian distribution, and the voiced-unvoiced switches which are essentially binary. To model these parameters simultaneously, we design the following energy function for GB-RBM:

$$\begin{aligned} E(\boldsymbol{v}^g, \boldsymbol{v}^b, \boldsymbol{h}) = &- \boldsymbol{h}^{\mathrm{T}}\boldsymbol{W}^g\boldsymbol{v}^g + \frac{1}{2}(\boldsymbol{v}^g - \boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{v}^g - \boldsymbol{\mu}) \\ &- \boldsymbol{h}^{\mathrm{T}}\boldsymbol{W}^b\boldsymbol{v}^b - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{v}^b - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{h}, \end{aligned} \tag{6}$$

where $\boldsymbol{v}^g$ and $\boldsymbol{v}^b$ are the Gaussian units and the Bernoulli units in the visible layer, $\boldsymbol{W}^g$ and $\boldsymbol{W}^b$ are the respective weight matrices, and $\boldsymbol{\mu}$ is the mean of $\boldsymbol{v}^g$. The conditional $\Pr(\boldsymbol{h}|\boldsymbol{v}^g, \boldsymbol{v}^b)$ can be similarly derived as:

$$\Pr(h_i = 1|\boldsymbol{v}^g, \boldsymbol{v}^b) = \sigma(\sum_j w_{ij}^g v_j^g + \sum_j w_{ij}^b v_j^b + a_i), \tag{7}$$

and $\Pr(v_j^b = 1|\boldsymbol{h})$ follows Eqn. (3). The conditional $\Pr(v_j^g|\boldsymbol{h})$ involves an integral over the continuous $\boldsymbol{v}^g$. We can show that:

$$\Pr(v_j^g|\boldsymbol{h}) = \mathcal{N}(v_j^g; \sum_i w_{ij}^g h_i + \boldsymbol{\mu}, 1). \tag{8}$$

Here we have assumed that the data is normalized to have unit variance. Given these defined conditional probabilities, CD-1 training is the same as described in Section 2.1.

2.3. Mixed Categorical-Bernoulli RBM (CB-RBM)

As mentioned previously that the posterior of hidden units can be inferred directly from an up-pass, to associate the inferred posteriors $\boldsymbol{v}^b$ with their corresponding indexed syllable label $\boldsymbol{l}^c$, we need to define the energy of the CB-RBM as follows:

$$\begin{aligned} E(\boldsymbol{l}^c, \boldsymbol{v}^b, \boldsymbol{h}) = &- \boldsymbol{h}^{\mathrm{T}}\boldsymbol{W}^l\boldsymbol{l}^c - \boldsymbol{b}^{c\mathrm{T}}\boldsymbol{l}^c \\ &- \boldsymbol{h}^{\mathrm{T}}\boldsymbol{W}^b\boldsymbol{v}^b - \boldsymbol{b}^{b\mathrm{T}}\boldsymbol{v}^b - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{h}. \end{aligned} \tag{9}$$

To make $\boldsymbol{l}^c \in \{0, 1\}^K$ follow a categorical distribution, i.e. representing the syllable identity using the 1-out-of-K code (K is the number of tonal syllables in the TTS system), we restrict $\boldsymbol{l}^c$ to have only K 1-out-of-K codes. It can be shown that $\Pr(l_j^c = 1|\boldsymbol{h})$ is defined by the soft-max:

$$\Pr(l_j^c = 1|\boldsymbol{h}) = \frac{\exp(\sum_i w_{ij}^c h_i + b_j^c)}{\sum_k \exp(\sum_i w_{ik}^c h_i + b_k^c)}. \tag{10}$$

The conditionals $\Pr(v_j^b = 1|\boldsymbol{h})$ and $\Pr(h_i = 1|\boldsymbol{l}^c, \boldsymbol{v}^b)$ take the same form as Eqn. (3) & (7) respectively in the CD-k training procedure.

3. DBN-BASED SPEECH SYNTHESIS

3.1. Training Stage

Given a syllable's start and end times, we extract 50 uniformly-spaced frames of 24-order Mel-Generalized Cepstrum coefficients (MGCs)[10] plus log-energy, 200 uniformly-spaced frames of voiced/unvoiced (V/UV) decisions and the corresponding log-F0 values within the syllable's boundary. The voiced and unvoiced frames are assigned 1s and 0s in their V/UV units, respectively. The log-F0 values for the unvoiced frames are set to be all 0 – a dummy value for log-F0. Both MGC and log-F0 have been normalized to have zero mean and unit variance.

The MGCs, log-F0 and V/UV units are concatenated to form a 1650-dimensional super-vector for each syllable, which is used as the visible layer $\boldsymbol{v}$ for GB-RBM. The posteriors of the hidden layer $\Pr(h_i^{(1)}|\boldsymbol{v}^g, \boldsymbol{v}^b)$ for all i yielded from a simple up-pass can be used as the visible data for training the immediate upper-layer B-RBM. Likewise, we stack up as many layers of B-RBMs as we want in a similar fashion. The joint distribution of the top hidden layer's posteriors (recursively propagated from below) and the associated syllable label is modeled by a CB-RBM.

3.2. Synthesis Stage

For an arbitrary text prompt, we look up the characters in a dictionary to find out their tonal syllable pronunciations. Starting from a clamped 1-out-of-K coded syllable label $\boldsymbol{l}^c$ and an initial all-zero $\boldsymbol{h}^{(N-1)}$ in the top-layer CB-RBM, we calculate the conditionals $\Pr(h_i^{(N)} = 1|\boldsymbol{l}^c, \boldsymbol{h}^{(N-1)})$ for all i in layer $\boldsymbol{h}^{(N)}$, $\Pr(h_j^{(N-1)} = 1|\boldsymbol{h}^{(N)})$ for all j in layer $\boldsymbol{h}^{(N-1)}$, in alternative fashion. This procedure continues until convergence or a maximum number of iterations is reached. $\Pr(h_j^{(N-1)}, \boldsymbol{l}^c)$ is then recursively passed down in the DBN to give out $\Pr(v_j^g|\boldsymbol{h}^{(1)})$ and $\Pr(v_j^b|\boldsymbol{h}^{(1)})$ for the Gaussian units and the Bernoulli units respectively in the GB-RBM at the bottom. The diagram of parameter generation from a DBN is shown in Fig. (1).

For each V/UV frame, we make a voicing decision depending on whether $\Pr(v_j^b|\boldsymbol{h}^{(1)}) > 0.5$. The log-F0 (if the corresponding V/UV decision is voiced) and MGC parameters are recovered via scaling by the standard deviation and offsetting by the mean. The duration of each syllable is the average estimated from the training data.

We apply a 25-point median filter on log-F0 trajectory to reduce noise. Then both MGCs and log-F0 parameter sequences are interpolated using cubic spline on the utterance level and decimated to yield parameter sequences with a constant 5-ms frame shift. The resulting speech parameter sequences are then fed to the Mel Log Spectral Approximation (MLSA) filter [11] for the final output signal.

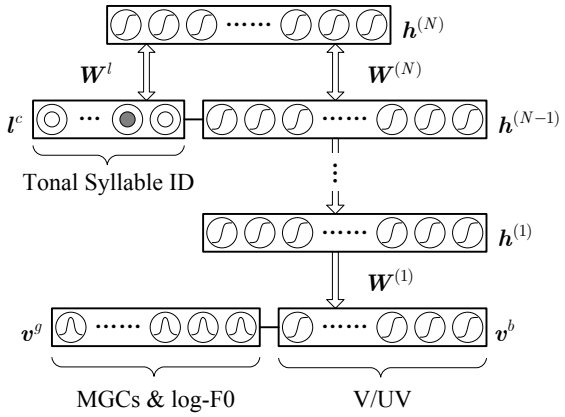

Fig. 1. Architecture of the MD-DBN for speech synthesis.

4. EXPERIMENTS

4.1. Experiment Setup

A manually transcribed Mandarin corpus recorded from a female speaker is used for the experiments. The training set contains 1,000 utterances with a total length of 80.9 minutes, including 23,727 syllable samples. All these samples are partitioned into 1,364 classes of tonal syllables. Another test set with 100 utterances is used for model architecture determination and the objective evaluation.

The RBMs are trained using stochastic gradient descent with a mini-batch size of 200 training samples. For GB-RBM, 400 epochs are executed with a learning rate of 0.01 while for B-RBMs and CB-RBM 200 epochs are executed with a learning rate of 0.1. During the weight updates, we apply a 0.9 momentum and a 0.001 weight decay.

The training procedure is accelerated by an NVIDIA Tesla M2090 GPU system. For an MD-DBN with 4 hidden layers and 2000 units per layer, the training takes about 1.1 hours. Each epoch of training GB-RBM, B-RBM and CB-RBM takes 3.5s, 3.7s and 5.7s respectively. A single GPU system runs at about 8 times faster than an 8-core 2.4 GHz Intel Xeon E5-2609 CPU system.

4.2. HMM-based Synthesis Baseline

We build a Mandarin HMM-based speech synthesis system on the same training set using a standard recipe [12]. MGCs and log-F0 together with their Δ and Δ^2 are modeled by multi-stream HMMs. Each syllable HMM has a left-to-right topology with 10 states. Initially, 416 mono-syllable HMMs are estimated as the seed for 1,364 tonal syllable HMMs. In the synthesis stage, speech parameters including MGCs and log-F0 are obtained by the maximum likelihood parameter generation algorithm [3], and are later used to produce the speech waveform through the MLSA filter. The syllable durations are the same as those in the DBN approach. For a

fair comparison, no contextual information or post-processing voice enhancement techniques are incorporated.

4.3. Optimizing MD-DBN Architecture

Finding an adequate architecture for MD-DBN is important in DBN-based speech synthesis. However, tweaking based on subjective evaluation is not practical due to the vast number of combinations of MD-DBN depth and layer width.

One commonly-used objective method in voice conversion as well as speech synthesis is to employ spectral distortion [13] between generated speech and target speech. Here we use Mel-Generalized Cepstral Distortion (MGCD) to determine the MD-DBN architecture. MGCD is the Euclidean distance between the MGCs of synthesized speech and that of original speech recording. All the test set prompts are synthesized to compute the average MGCDs for HMM baseline and MD-DBN with different number of hidden layers and different number of units in each hidden layer. As the syllable duration of the speech from the test set and that of HMM and MD-DBN can be different, we align the MGCs according to the syllable boundary.

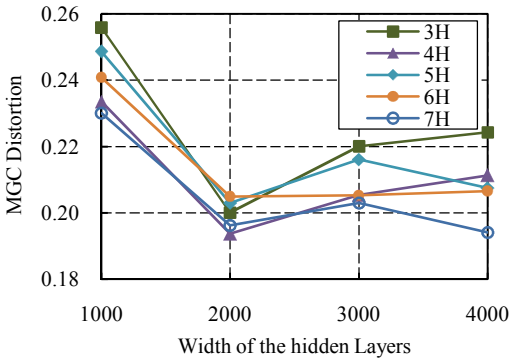

Fig. 2. MGCD as a function of the width of the hidden layers. 3H-7H: No. of hidden layers.

The MGCD of the HMM baseline is 0.223, while the DBN approach archives better result with a minimal MGCD of 0.194. As shown in Fig. (2), the MGC distortions are high when the layer width is 1,000, and too many units (4,000) in the hidden layer causes large variance of MGCD. The MD-DBN with 4 hidden layers and hidden layer width of 2,000 units consistently gives the best MGCD. Hence we will be using these parameters for the remaining of the evaluation.

4.4. Subjective Evaluation

A Mean Opinion Scoring (MOS) test is conducted to compare the subjective perception between the DBN approach and the HMM baseline, together with a hybrid approach (MIX) using MGCs from DBN and log-F0 from HMM. In the MOS test, each of the 10 experienced listeners is asked to rate 10 utterances synthesized by the DBN and the HMM using a 5-point scale (5:excellent, 4:good, 3:fair, 2:poor, 1:bad). The MOS result is shown in Table (1).

System	MOS
HMM	2.86
DBN	2.88
MIX: DBN MGCs + HMM Log-F0	3.09

Table 1. MOS test result.

Although the overall MOS score shows a draw between the DBN and the HMM, the two sets of synthesized speech sounds different[1]. Without post-processing, HMM's voice sounds quite muffled, but the prosody remains smooth and stable. The DBN voice sounds much clearer than the HMM baseline, and the prosody is lively. However, it seems that tonal and U/VU decision errors have a relatively higher chance to occur in the DBN approach, which can be the main reason that lowers its MOS score. The MIX approach gets highest score, which probably suggests that the spectrum can be captured by DBN appropriately but HMM does a better job in F0 modeling.

The cross-comparison reveals more details. With the same log-F0 curve (HMM vs. MIX), MGCs generated by DBN leads to higher MOS scores. This result agrees with the objective MGCD measure ($\text{MGCD}_{\text{HMM}} > \text{MGCD}_{\text{DBN}}$). With the same MGCs (DBN vs. MIX), log-F0 curve generated by HMM results in higher MOS scores. It can be seen that listeners prefer smoother F0 patterns when there is no higher-level prosody control of the synthesized speech.

5. CONCLUSIONS AND FUTURE WORK

We have described a DBN model with a multi-distribution visible layer for statistical parametric speech synthesis, in which the spectrum and F0 are modeled simultaneously in a unified framework. It is shown that DBN models spectrum better than HMM and achieves an overall performance that is comparable to context-independent HMM synthesis. Future directions for improvement include: (1) introducing context information to improve the prosody; and (2) better modeling of the F0 contour with higher resolution of frame sampling.

6. ACKNOWLEDGMENT

The authors would like to thank Dr. Li Deng from MSR and Prof. Fei Sha from USC for fruitful exchanges.

[1]The synthesized speech samples can be downloaded from `http://www.se.cuhk.edu.hk/~sykang/icassp2013/`

7. REFERENCES

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Eurospeech*, 1999, pp. 2347–2350.

[2] K. Tokuda, T. Mausko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. & Syst.*, vol. E85-D, pp. 455–464, 2002.

[3] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *ICASSP*, 2000, pp. 1315–1318.

[4] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[5] Joshua M Susskind, Geoffrey E Hinton, Javier R Movellan, and Adam K Anderson, "Generating facial expressions with deep belief nets," *Affective Computing, Emotion Modelling, Synthesis and Recognition*, pp. 421–440, 2008.

[6] G. W. Taylor, G. E. Hinton, and S. T. Roweis, "Two distributed-state models for generating high-dimensional time series," *Journal of Machine Learning Research*, vol. 12, pp. 1025–1068, 2011.

[7] George E Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.

[8] A.W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *ICASSP*, 2007, pp. 1229–1232.

[9] A. Mohamed, G. E. Dahl, and G. E. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[10] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," in *ICSLP*, 1994.

[11] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *ICASSP*, 1992, pp. 137–140.

[12] Z. Shuang, S. Kang, Q. Shi, Y. Qin, and L. Cai, "Syllable HMM based mandarin TTS and comparison with concatenative TTS," in *INTERSPEECH*, 2009, pp. 1767–1770.

[13] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *ICASSP*, 2009, pp. 3893–3896.