DEEP LEARNING WITH MAXIMAL FIGURE-OF-MERIT COST TO ADVANCE MULTI-LABEL SPEECH ATTRIBUTE DETECTION.

Ivan Kukanov[1] *Ville Hautamäki*[1] *Sabato Marco Siniscalchi*[2,3] *Kehuang Li*[3]

[1]School of Computing, University of Eastern Finland, Finland
[2]Faculty of Architecture and Engineering, University of Enna "Kore", Italy
[3]School of ECE, Georgia Institute of Technology, USA

ABSTRACT

In this work, we are interested in boosting speech attribute detection by formulating it as a multi-label classification task, and deep neural networks (DNNs) are used to design speech attribute detectors. A straightforward way to tackle the speech attribute detection task is to estimate DNN parameters using the mean squared error (MSE) loss function and employ a sigmoid function in the DNN output nodes. A more principled way is nonetheless to incorporate the micro-F1 measure, which is a widely used metric in the multi-label classification, into the DNN loss function to directly improve the metric of interest at training time. Micro-F1 is not differentiable, yet we overcome such a problem by casting our task under the maximal figure-of-merit (MFoM) learning framework. The results demonstrate that our MFoM approach consistently outperforms the baseline systems.

Index Terms— Speech articulatory attributes detection, deep neural networks, convolutional neural networks, maximal figure-of-merit, foreign accent recognition

1. INTRODUCTION

In the past, several studies have suggested that a proper integration of some knowledge sources into standard ASR systems may be beneficial. For example, accent, gender, and wide-phonetic knowledge were incorporated into the acoustic modeling design of an HMM-based system in [1] with good results. Speech knowledge represented by phonetically motivated acoustic parameters [2] and articulatory-motivated distinctive features [3, 4] have been embedded into an HMM-based recognizer at the front-end. In [3], a set of classifiers learns the mapping between the acoustic space and the distinctive features space. The outputs of these classifiers are combined with the cepstral vector to form an extended front-end which is then used to train the HMM-based acoustic models. This ASR system achieves better performance than the conventional cepstral-based system. In [2], an extended front-end is created by appending some acoustic parameters to the cepstral vector. In [5, 6, 7], articulatory knowledge is integrated at the HMM state level. In [5], a set of ANNs is used to score articulatory-motivated features for manner and place of articulation. The posterior feature probability outputs from each network are combined by a higher-level integrative ANN which maps them to phone probabilities. This ANN is used as an emission probability estimator in the HMM framework. It is shown that these new features improve robustness against noise at low signal-to-noise ratios; moreover, several methods for system combination are outlined. In [6], a stream architecture was proposed to augment acoustic models based on context-dependent sub-words with articulatory-motivated acoustic models. Automatic speech attribute transcription (ASAT) [8, 9] is a bottom-up framework that first detects a collection of speech attribute cues and then integrates such cues to make linguistic validations. A typical ASAT system uses the articulatory-based phonological features studied earlier, e.g., [3, 10], in a new detection-based framework. Several successful applications of ASAT framework have been proposed in different domains of speech processing, such as phoneme recognition [11], foreign accent recognition [12], language recognition [13].

A critical yet fundamental component of those above mentioned knowledge-based approaches, especially ASAT, is to build a set of data-driven models to reliably detect a collection of speech attribute cues, such as manner and place of articulation. In this paper, we thus aim to extended previous work presented in [12] and enhance attribute detection. To this end, we cast the problem of the attribute extraction from a short-term spectral representation of the speech signal into a *multi-label classification* problem [14, 15]. According to the *multi-label learning* theory [16], each observation can be associated with multiple labels (*a.k.a.* attribute classes) at the same time. Instead of classification error rate the micro-F1 metric is often adopted as the accuracy measure for a multi-label classifier. Here we explore two multi-label approaches. The first approach models all attribute detectors with single DNN (jointly using the whole dataset), where each output neuron, having a sigmoid activation function, is associated with a single attribute class and produces a confidence score independently of the other output neurons. Moreover, the *mean squared error* (MSE) objective function is used to es-

timate the DNN parameters. We refer to this as the *baseline* approach. The second approach explores the *maximal figure-of-merit* (MFoM) [17, 18] learning idea. It allows embedding the micro-F1 metric directly into the loss function of the DNN and makes it possible to directly improve the measure of interest. MFoM tries to improve the *decision boundary* [17] using the output sigmoid scores without needing any intermediate calibration. Moreover, we use the training set label information about the joint attribute classes that have "several-hot-labels". Such multi-label information is embedded into the *misclassification distance measure* and defines the strategies for the discriminative functions either "one-vs-others" or "units-vs-zeros", these strategies define the classes that are competing and that are cooperating.

We archive the consistent improvements with the MFoM-micro-F1 approach and "units-vs-zeros" misclassification measure for the detectors. The most significant result is attained for the place of articulation detectors by the fusion DNN system with micro-F1 error 31.86% versus the baseline system with 33.35%.

2. SPEECH ATTRIBUTE MODELING

Let us define a training set as $\mathbb{T} = \left\{ (\mathbf{x}_i, \mathbf{y}_i) | \, i = \overline{1,t} \right\}$, i.e. t pairs of training sample $\mathbf{x}_i \in \mathbb{R}^{\mathbf{D}}$ of dimension D and binary vector of labels $\mathbf{y}_i \in \{0,1\}^M$. Suppose we have M classes $\mathbb{C} = \{C_k | k = \overline{1,M}\}$; then in multi-label multi-class classification problem, the binary vector of labels $\mathbf{y}_i$ has several unit marks (several-hot labels), assigning sample $\mathbf{x}_i$ to several classes at the same time. Every data point $\mathbf{x}_i$ must be assigned to at least one class C_k, thus all data points' set $\mathbb{X} \subset \bigcup_{k=1}^{M} C_k$ is less than the total size of all subsets C_k, because in the multi-label case the classes intersect.

The goal of the learning system is to train a multi-label classifier $\mathbf{H} : \mathbb{X} \to 2^{|\mathbb{Y}|}$, which is able to assign a subset of labels to any sample. In practice, some systems do not produce multiple-labels directly, but emit real-valued confidence score $\mathbf{G} : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ that sample $\mathbf{x}_i$ is labeled with $\mathbf{y}_i$, where function $\mathbf{G}$ is called as *discriminative function*. Therefore, if the number of classes is M, then multi-label detectors can associate a sample with 2^M labels, so the search space is much more diverse than in the single-label case.

Two types of speech articulatory attribute detectors are modeled, these are manner (7 classes, $M = 7$) and place (11 classes, $M = 11$) [19]. In the *baseline system*, we model both types of attributes with their own DNNs. Two topologies of DNN are explored: the deep neural network with layer-wise RBM pre-training (DBN-DNN) and 1D convolutional neural network followed by the fully connected layers (1D CNN).

During the training phase, the context-dependent speech frame vector $\mathbf{x}_i$ is fitted in the input of the DNNs, and its associated target label $\mathbf{y}_i$ is represented as a binary vector, whose dimension is equal to the number of attribute classes (7 or 11). The MSE loss function is used to optimize the network parameters' set $\mathbb{W} = \left\{ \mathbf{W}_n | \, n = \overline{1, L-1} \right\}$ for a network with L layers.

It is supposed that, given an instance $\mathbf{x}$ and its "several-hot-labels" binary vector $\mathbf{y}$, a successful learning system will tend to output larger score values for labels in $\mathbf{y}$ than those not in $\mathbf{y}$; this is formulated in terms of the *discriminative function* [16]. Assume for each class C_k the discriminative function is defined as $g_k(\mathbf{x}; \mathbb{W})$, then the predicted class for any sample $\mathbf{x}$ is $C_y^* = \operatorname{argmax}_y g_y(\mathbf{x}; \mathbb{W})$ [20]. In the multi-label case a threshold is applied in order to chose several candidates. In the baseline DNN system, let us suppose that the discriminative function for an individual class C_k is sigmoid output score

$$g_k(\mathbf{z}) = \sigma_k(\mathbf{z}), \quad k = \overline{1, M}, \tag{1}$$

where $\mathbf{z} = \mathbf{W}_{L-1}\mathbf{x}^{L-1}$ is the vector $\mathbf{z} = (z_1, \ldots, z_M)^T$ of the pre-activation values of the last network layer L, i.e. before feeding $\mathbf{z}$ to the sigmoid activations of the last network layer, $\mathbf{x}^{L-1}$ is a sample $\mathbf{x}$ forwarded through the network up to the $L-1$ layer. In order to reveal the detected classes we apply a thresholding decision rule to the sigmoid discriminative functions for the baseline deep network

$$C_k^* = \mathbf{1}\left(\sigma_k(\mathbf{z}) > t\right), \quad k = \overline{1, M}, \tag{2}$$

where the $\mathbf{1}(\cdot)$ indicator function means if the sigmoid is larger than threshold t, then a sample belongs to the class C_k^* and labeled as 1, otherwise the sample is labeled as 0. In the experiments, we threshold the scores with $t = 0.5$.

The micro-average F_1^μ metric is usually exploited in order to measure a multi-label classification performance [21] over M classes. By definition, the micro-F1 measure is the harmonic mean of precision P^μ and recall R^μ

$$F_1^\mu = \frac{2 \cdot P^\mu \cdot R^\mu}{P^\mu + R^\mu} = \frac{2 \cdot \sum\limits_{k=1}^{M} TP_k}{\sum\limits_{k=1}^{M} (TP_k + 2 \cdot FP_k + FN_k)}, \tag{3}$$

and can be expressed as a function of counts of true positives TP_k, false positives FP_k and false negatives FN_k.

3. MFOM APPROACH ON NEURAL NETWORKS

MFoM learning approach is the heart of the *generalized probabilistic descent method* [22]. We propose network architecture with MFoM which is shown in Figure 1. Here we formalize the inference of the MFoM for multi-label performance measure for micro F_1^μ objective function. We then optimize it as the loss function of the DNN. First, the *misclassification measure* [18] for each class C_k is defined as

$$d_k(\mathbf{z}) = -g_k(\mathbf{z}) + \ln\left(\frac{1}{M-1} \sum_{\substack{j=1 \\ j \neq k}}^{M} e^{\eta g_j(\mathbf{z})} \right), \tag{4}$$

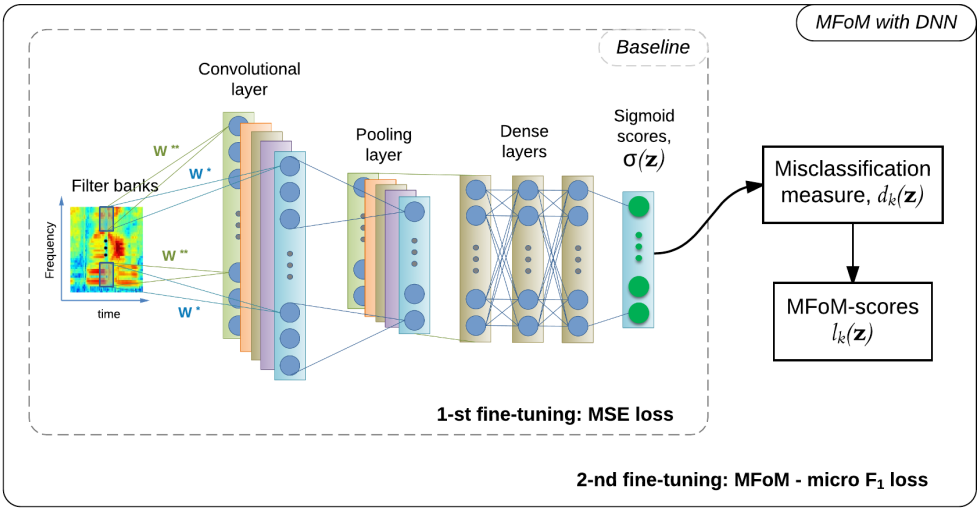

Fig. 1. Multi-label architecture with 1D convolutional neural network (1D CNN) is trained in two phases. We train the initial network with the MSE loss function; the output scores are sigmoids $\sigma(\mathbf{z})$. Then we fine-tune the network transforming the sigmoid scores into MFoM scores and optimizing the MFoM-micro-F_1 loss.

where $g_k(\mathbf{z})$ is the discriminative function, η is a smooth positive real-valued constant. The left-side term of the (4) is called the target model and the right-side is the geometrical mean of the competing models. The misclassification measure defines the distance between the target and its decision surface. Varying the parameter η enables the simulation of various decision rules and in the extreme case when $\eta \to +\infty$, the geometrical average is becoming a *maximum metric* [22], i.e., converges to the highest score among the competing classes. The sign of the misclassification measure indicates the correctness of classification: $d_k(\cdot) < 0$ indicates the predicted class is correct, and vice versa. The absolute value of the d_k quantifies the separation between the correct and competing classes [18].

In the case of optimizing MFoM for deep learning, we decide to use the same discrimination function (1) as in the baseline approach. After some reorganization and sigmoid substitution, the misclassification function is obtained

$$d_k(\mathbf{z}) = \frac{1}{\eta} \ln \left[\frac{1}{M-1} \left(\frac{1}{s\sigma_k(\mathbf{z})} - 1 \right) \right], \tag{5}$$

where for simplicity of notation we define

$$s\sigma_k(\mathbf{z}) \triangleq \frac{e^{\eta \sigma_k(\mathbf{z})}}{\sum_{j=1}^{M} e^{\eta \sigma_j(\mathbf{z})}}. \tag{6}$$

which is seen as the softmax function.

The misclassification measure can be thought as the "strategy function", because it defines the multi-label decision rule, cooperating and competing models. In the model equation (4) the rough decision takes place where strategy **"one-vs-others"** works, i.e. target model competes the same time with all others (anti-target models). Here we propose to define the alternative misclassification measure, which is called as **"units-vs-zeros"**. It explicitly incorporates the label information into the misclassification measure. It means that for current class C_k labeled as 1, the competing models C_j will be considered only these with labels 0 and vice versa, if C_k is labeled as 0. Then we can reformulate the misclassification measure (4)

$$d_k(\mathbf{z}) = -g_k(\mathbf{z}) + \frac{1}{\eta} \ln \left(\frac{1}{|\mathbf{I}|} \sum_{j \in \mathbf{I}} e^{\eta g_j(\mathbf{z})} \right), \tag{7}$$

$$\begin{cases} \text{if } C_k \text{ is } 1 \Rightarrow \mathbf{I} = \mathbf{y}_{\{0\}}, \\ \text{if } C_k \text{ is } 0 \Rightarrow \mathbf{I} = \mathbf{y}_{\{1\}}, \end{cases} \tag{8}$$

where, for current sample $\mathbf{x}$ and its label $\mathbf{y}$, $\mathbf{I}$ is an index set, $\mathbf{y}_{\{1\}}$ is the set of unit indexes and $\mathbf{y}_{\{0\}}$ is the set of zero indexes in the label vector $\mathbf{y}$. Thus we directly choose the competing anti-models using label information.

Further, the *class loss* function (sigmoid function) is introduced to estimate the count of misclassified samples; it is a smooth version of the error step function [22]

$$l_k(\mathbf{z}) = \frac{1}{1 + \exp\left[-\alpha_k d_k(\mathbf{z}) - \beta\right]}, \tag{9}$$

where α_k and β are two positive parameters controlling learning speed, for more information about these tuning see [17]. It should be close to 0 for correct detection and 1 for incorrect.

Therefore, the MFoM framework for the network objective function consists of three key objects: 1) discriminative function (1), which is sigmoid activation units of the last layer of the network, 2) misclassification measure (4) or (7), and 3) smoothed class loss function (9). Now that these components are introduced, we can express the cost micro F_1^μ function for the neural networks. The discrete F_1^μ measure for the multi-class case of M classes [15]

$$F_1^\mu = \frac{2\sum_{k=1}^{M} TP_k}{\sum_{k=1}^{M} FP_k + \sum_{k=1}^{M} TP_k + \sum_{k=1}^{M} |C_k|}, \tag{10}$$

where the number of samples in class C_k

$$|C_k| = TP_k + FN_k. \tag{11}$$

The smooth approximation of the error counts of the true positive and false positive[17]

$$TP_k \approx (1 - l_k(\mathbf{z}_i)) \cdot \mathbf{1}(\mathbf{x}_i \in C_k), \tag{12}$$

$$FP_k \approx (1 - l_k(\mathbf{z}_i)) \cdot \mathbf{1}(\mathbf{x}_i \notin C_k), \tag{13}$$

where $\mathbf{1}(\cdot)$ is the indicator function of the logical expression. Thus, the differentiable F_1^μ is presented. Our task is to minimize the objective function during training the neural network optimizing its parameters

$$E = 1 - F_1^\mu \to \min_{\mathbf{W}}. \tag{14}$$

The derivative of the objective function on the mini-batch $\mathbf{T}$ using smoothed TP_k and FP_k is derived

$$\nabla E(\mathbf{X}, \mathbf{W}) = A \cdot (\omega_1 \cdot \Delta FN + \omega_2 \Delta FP)$$

where

$$A = \frac{2}{(FP + TP + |C|)^2},$$

$$\omega_1 = |C| + FP, \quad \omega_2 = TP,$$

$$\Delta FN = \sum_{\mathbf{x} \in T} \sum_{k=1}^{M} \frac{\partial l_k(\mathbf{z})}{\partial \mathbf{z}} \cdot \mathbf{1}(\mathbf{x} \in C_k),$$

and

$$\Delta FP = -\sum_{\mathbf{x} \in T} \sum_{k=1}^{M} \frac{\partial l_k(\mathbf{z})}{\partial \mathbf{z}} \cdot \mathbf{1}(\mathbf{x} \notin C_k).$$

The Jacobian matrix of the class loss function is

$$J(\mathbf{z}) = \frac{\partial l_k(\mathbf{z})}{\partial \mathbf{z}} = \begin{cases} -l'_k \cdot \sigma'_k, \text{ where } m = k \\ \frac{l'_k \cdot s\sigma_m(\mathbf{z})}{1 - s\sigma_k} \cdot \sigma'_m, \text{ where } m \neq k \end{cases}$$

where $\sigma'_k = \sigma_k(\mathbf{z})(1 - \sigma_k(\mathbf{z}))$ is the derivative of the neuron activation function and $l'_k = \alpha_k l_k(1 - l_k)$.

We should notice here, when the misclassification measure $d_k(\cdot) = 0$ or close to zero, it means that the target sample is close to the decision surface and there is uncertainty for the discrimination of that sample, which plays important role in learning the classifier [22]. At the same time the class loss function $l_k(\cdot) = 0.5$ reaches the maximal value of the $l'_k = \alpha_k l_k(1 - l_k)$.

Table 1. *The networks have from 1 to 10 hidden dense layers. We select the best baseline topologies of the networks for the manner of articulation.*

Topology	# of units	# of hid. layers	F_1^μ error, %
CNN	64	6	13.59
CNN	128	3	**13.42**
DBN-DNN	512	2	16.05
DBN-DNN	1024	2	15.79

Table 2. *The same as in Table 1, but for the place of articulation.*

Topology	# of units	# of hid. layers	F_1^μ error, %
CNN	64	8	34.46
CNN	128	7	**34.11**
DBN-DNN	512	4	38.00
DBN-DNN	1024	5	37.18

Table 3. *The best baseline performance for the fusion DNN topologies.*

Topology	units	# of hid. layers	F_1^μ error, %	
			manner	place
CNN	64	4	13.64	34.21
CNN	128	5	**13.40**	**33.35**
DBN-DNN	512	2	16.37	41.79
DBN-DNN	1024	2	16.09	40.94

4. EXPERIMENTS

Attribute detection experiments are conducted on the "stories" part of the OGI Multi-language Telephone Speech corpus [23]. This corpus has manual phonetic transcriptions for six languages: English, German, Hindi, Japanese, Mandarin, and Spanish. Data from each language were pooled together to obtain 5.57 hours of training and 0.52 hours of validation data. Attribute ground-truth is produced using mapping phonological tables from phonemes into place and manner articulatory attributes [19]. One phoneme can be mapped into several attributes, thus one observation can have several labels[19].

During the baseline experiment, we train two types of networks; these are deep neural networks with layer-wise pair-wise pre-training with restricted Boltzmann machine and 1D convolutional neural network with two convolutional and pooling layers, see Figure 1. These two architectures with different hyperparameters we train for each attribute type (manner and place). Also, we train a fusion network, whose output layer jointly emits scores for both manner and place attributes. The mean-squared error objective function is optimized and the output detection scores are sigmoid scores $\sigma(\mathbf{z})$ of the last layer of the networks, see Figure 1. After applying the decision rule (2), with threshold $t = 0.5$, to the sigmoid scores, we calculate the performance of the system with the discrete micro-F1 measure (10). The best settings for number of layers and neurons are presented in Tables 1-3.

As the starting point for the proposed MFoM-micro-F1 approach, we explore the networks trained with the MSE, which show the highest performance on the baseline (bold numbers in Tables 1-3). The OGI training dataset is forwarded through the MSE trained networks, then the output sigmoid scores are turned into MFoM scores with misclassification measure (4) or (7) and class loss function (9), as shown in Figure 1. The micro-F1 objective function (14) is optimized.

4.1. DBN-DNN Topology

The input feature vector for DBN-DNN is a 45-dimension mean-normalized log-filter bank features with up to second-order derivatives and a context window of 11 frames, forming an input vector of 495-dimension (45×11). The number of output classes is equal to 7 for manner and 11 for place, or 18 for fusion. Indeed, the *"other"* output class is added to both DNN to handle possible unlabelled frames. DNN topologies with 512 and 1024 number of units were studied, the number of hidden layers is varying from 1 up to 10.

All DBN-DNN topologies are initialized with the stacked restricted Boltzmann machines using layer by layer generative pre-training. The pre-training algorithm is contrastive divergence with 1-step of Markov Chain Monte Carlo sampling (CD-1). The first RBM has Gaussian-Bernoulli units and trained with the initial learning rate of 0.01. The following RBMs have Bernoulli-Bernoulli units and a learning rate of 0.4.

After pre-training the weights and stacking all the layers, the final output sigmoid layer was concatenated with the DNN. The fine-tuning for the final weights training is done by mini-batch stochastic gradient descent with learning rate of 0.008. Mini-batch size is equal to 128 observations. The Mean Square Error objective function was optimized during fine-tuning. All units in the DNN have sigmoid activation function. Notice, all settings and parameters for the DNNs are conventional in speech community [24].

Table 4. *The MFoM micro-F1 error (%) with **"one-vs-others"** and **"units-vs-zeros"** misclassification measures for manner, place and fusion systems.*

Detectors	"one-vs-others"	"units-vs-zeros"
Manner	13.92	13.35
Place	37.22	32.43
Fuse-Manner	14.23	13.29
Fuse-Place	36.75	31.86

4.2. CNN Topology

In this work a 1D CNN topology [25] has been explored, see Figure 1. It takes the *input feature maps*, which are organized from 40-dimensional log-Mel filter bank features, their first and second-order derivatives and context window is size of 11 frames. In total $3 \times 11 = 33$ input feature maps to which we apply 1D convolution mapping along the frequency axis. The CNN consists of a convolutional layer and max-pooling layer, then from 1 up to 10 fully-connected hidden layers each of them has 64 or 128 sigmoid units. Similar to the DBN-DNN settings, in the case of the CNN, we optimize the MSE cost function. Output layer has sigmoid units and produces sigmoid confidence scores for every attribute (7 for the manner and 11 for the place, or 18 for the fusion) per each speech frame.

Convolutional layer in the CNN has 128 feature maps, each of which has the size of 33 frequency bands, i.e. these are produced by convolving each input feature map with the filter size of 8 ($1 + (40 - 8) = 33$). After that max-pooling layer outputs the maximum values over a non-overlapping window covering the outputs of every three frequency bands in each feature map (i.e. pooling size is 3), down-sampling the overall convolutional layer output to three times smaller. Then the output of the max-pooling layer is fed to the fully-connected feed-forward part of the CNN. The fully-connected layers are pre-trained using RBMs as described in [25].

5. RESULTS

In the baseline experiments, we train the networks with different settings of hyperparameters. Number of dense hidden layers varies from 1 to 10; every hidden layer has 64 neurons or 128 neurons for 1D CNN topology and 512 neurons or 1024 neurons for DBN-DNN topology. In the Tables 1-3, we summarize the best baseline topologies of the networks. We notice the tendency that CNN topologies dramatically outperform the DBN-DNNs. The CNN topology with 128 neurons in its hidden dense layers versus DBN-DNN with 1024 neurons improves detection of the manner attributes from 15.79% to 13.42% and the place from 37.18% to 34.11%. It is interesting to note that after the fusion manner and place attributes into the joint network, we increase the performance

with CNN topology from 13.42% to 13.40% for the manner and from 34.11% to 33.35% for the place attributes. In opposite, the DBN-DNN increase the detection error from 15.79% to 16.09% for the manner and from 37.18% to 40.94% for the place attributes.

Approach with the maximal figure-of-merit (MFoM) learning framework using the proposed "units-vs-zeros" misclassification measure significantly outperforms "one-vs-others" training strategy, see Table 4. On top of that, "units-vs-zeros" strategy improves the baseline performance, especially on the fusion system for the manner from 13.42% to 13.29% and for the place part from 34.11% to 31.86%.

6. CONCLUSIONS

In this paper, we have investigated the use of deep architectures to improve the articulatory attribute detectors, namely manner and place. In particular, we have designed two deep neural networks DBN-DNN and 1D CNN. On the baseline settings with sigmoid outputs and the MSE loss function, the best result is shown by 1D CNN. In the baseline approach, we can notice that fusion system slightly improves the manner and the place detectors especially on CNN with 128 units in fully connected part. It improves place of articulation detectors from 34.11% to 33.35%. Moreover, we propose the new approach using the maximal figure-of-merit (MFoM) learning framework. The MFoM is used as the optimization approach which allows to directly incorporate micro-F1 multi-label classification metric into DNN training loss function. Also, we utilize the training set label information about the joint attribute classes which have "several-hot-labels". Such multi-label information is embedded into the misclassification distance measure and defines the strategies for the discriminative functions either "one-vs-others" or "units-vs-zeros", these strategies define the classes which are competing and which are cooperating. Experimental results have demonstrated that the "units-vs-zeros" strategy significantly improved the detection performance, especially on the fusion system for the manner from 13.42% to 13.29% and for the place part from 34.11% to 31.86%. We intend to expand further this line of research by exploiting multi-task learning approach and recurrent deep neural networks.

7. REFERENCES

[1] S. Sakti, K. Markov, and S. Nakamura, "An HMM acoustic model incorporating various additional knowledge sources," in *Proc. Interspeech*, Antwerp, Belgium, Sept. 2007, pp. 2117–2120.

[2] N. N. Bitar and C. Y. Espy-Wilson, "Knowledge-based parameters for HMM speech recognition," in *Proc. ICASSP*, Atlanta, USA, May 1996, pp. 29–32.

[3] E. Eide, "Distinctive features for use in an automatic speech recognition system," in *Proc. EuroSpeech*, Aalborg, Denmark, Sept. 2001, pp. 1613–1616.

[4] B. Launay, O. Siohan, A. C. Surendran, and C.-H. Lee, "Towards knowledge-based features for HMM based large vocabulary automatic speech recognition," in *Proc. ICASSP*, Orlando, USA, May 2002, pp. 817–820.

[5] K. Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments," in *Proc. ICSLP*, Sydney, Australia, Nov./Dec. 1998, pp. 891–894.

[6] F. Metze and A. Waibel, "A flexible stream architecture for ASR using articulatory features," in *Proc. ICSLP*, Denver, USA, Sept. 2002, pp. 16–20.

[7] M. Richardson, J. Blimes, and C. Diorio, "Hidden articulator Markov models for speech recognition," *Speech Communication*, vol. 41, no. 2, pp. 511–529, 2003.

[8] C.-H. Lee and S. M. Siniscalchi, "An information-extraction approach to speech processing: Analysis, detection, verification, and recognition," *Proceedings of the IEEE*, vol. 101(5), pp. 1089–1115, 2013.

[9] I. Bromberg, Q. Qian, J. Hou, J. Li, C. Ma, B. Matthews, A. Moreno-Daniel, J. Morris, S. M. Siniscalchi, Y. Tsao, and Y. Wang, "Detection-based ASR in the automatic speech attribute transcription project," in *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium*, pp. 1829–1832.

[10] L. Deng and D. Sun, "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features," *J. Acoust. Soc. Am.*, vol. 85, no. 5, pp. 2702–2719, 1994.

[11] S. M. Siniscalchi, P. Schwarz, and C.-H. Lee, "High-accuracy phone recognition by combining high-performance lattice generation and knowledge based rescoring," in *Proc. ICASSP*, Honolulu, HI, USA, Apr. 2007, pp. IV–869–V–872.

[12] V. Hautamäki, S. M. Siniscalchi, H. Behravan, V. M. Salerno, and I. Kukanov, "Boosting universal speech attributes classification with deep neural network for foreign accent characterization," *INTERSPEECH*, 2015.

[13] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Universal attribute characterization of spoken languages for automatic spoken language recognition," *Computer Speech & Language*, vol. 27, no. 1, pp. 209–227, 2013.

[14] J. Nam, J. Kim, E. Loza Mencía, I. Gurevych, and J. Fürnkranz, "Large-scale multi-label text classification - revisiting neural networks," in *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD-14), Part 2*, ser. Lecture Notes in Computer Science, T. Calders, F. Esposito, E. Hüllermeier, and R. Meo, Eds. Springer Berlin Heidelberg, 2014, vol. 8725, pp. 437–452.

[15] M.-L. Zhang, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1338–1351, Oct. 2006.

[16] M. S. Sorower, "A literature survey on algorithms for multi-label learning," Tech. Rep., 2010.

[17] S. Gao, W. Wu, C.-H. Lee, and T.-S. Chua, "A maximal figure-of-merit (MFoM)-learning approach to robust classifier design for text categorization," *ACM Transactions on Information Systems*, vol. 24, p. 2006, 2006.

[18] K. Li, Z. Huang, Y.-C. Cheng, and C.-H. Lee, "A maximal figure-of-merit learning approach to maximizing mean average precision with deep neural network based classifiers," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Institute of Electrical & Electronics Engineers (IEEE), may 2014.

[19] P. Gasiorowski and R. Lew, "Francis katamba, an introduction to phonology." *Journal of the International Phonetic Association*, vol. 22, no. 1-2, p. 70, jun 1992.

[20] C. M. Bishop, *Pattern Recognition and Machine Learning*, M. Jordan, J. Kleinberg, and B. Scholkopf, Eds. Springer, 2006.

[21] M. Zhang and Z. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. PP, no. 99, p. 1, 2013.

[22] Y. H. Hu, *Handbook of Neural Network Signal Processing*, 1st ed., J.-N. Hwang and J.-N. Hwang, Eds. Boca Raton, FL, USA: CRC Press, Inc., 2000.

[23] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI Multi-Language telephone speech corpus," in *Proceedings of the International Conference on Spoken Language Processing*, 1992, pp. 895–898.

[24] G. Hinton, L. Deng, D. Yu, G. Dahl, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, nov 2012.

[25] T. N. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, 2013, pp. 8614–8618.