This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2964953, IEEE/ACM Transactions on Audio, Speech, and Language Processing



# Maximal Figure-of-Merit Framework to Detect Multi-label Phonetic Features for Spoken Language Recognition

Ivan Kukanov, Trung Ngo Trong, Ville Hautamäki, *Member, IEEE,* Sabato Marco Siniscalchi, *Senior Member, IEEE,* Valerio Mario Salerno, Kong Aik Lee, *Senior Member, IEEE*

***Abstract*—Bottleneck features (BNFs) generated with a deep neural network (DNN) have proven to boost spoken language recognition accuracy over basic spectral features significantly. However, BNFs are commonly extracted using language-dependent tied-context phone states as learning targets. Moreover, BNFs are less phonetically expressive than the output layer in a DNN, which is usually not used as a speech feature because of its very high dimensionality hindering further post-processing. In this work, we put forth a novel deep learning framework to overcome all of the above issues and evaluate it on the 2017 NIST Language Recognition Evaluation (LRE) challenge. We use manner and place of articulation as speech attributes, which lead to low-dimensional "universal" phonetic features that can be defined across all spoken languages. To model the asynchronous nature of the speech attributes while capturing their intrinsic relationships in a given speech segment, we introduce a new training scheme for deep architectures based on a Maximal Figure of Merit (MFoM) objective. MFoM introduces non-differentiable metrics into the backpropagation-based approach, which is elegantly solved in the proposed framework. The experimental evidence collected on the recent NIST LRE 2017 challenge demonstrates the effectiveness of our solution. In fact, the performance of speech language recognition (SLR) systems based on spectral features is improved for more than 5% absolute Cavg. Finally, the F1 metric can be brought from 77.6% up to 78.1% by combining the conventional baseline phonetic BNFs with the proposed articulatory attribute features.**

***Index Terms*—Convolutional recurrent neural network, speech articulatory attributes, maximal figure-of-merit, deep bottleneck features, spoken language recognition.**

## I. Introduction

WE can recognize a written language by analyzing its $n$-gram distribution, where an $n$-gram is a sequence of $n$ words, and that has been known since the time of Shannon [1], at least. It was, therefore, natural to extend that idea to the automatic spoken language recognition (SLR) task [2], where a *language model* of the automatic speech recognition (ASR) output was fed to a classifier, such as a support vector machine (SVM) [3], to perform language classification. This approach is commonly referred to as *token-based* [4], and it is also known as the *phonotactic* approach [5] if the ASR output is used to obtain tokens.

Another approach to language recognition is the *spectral* approach, in which short-term spectral magnitude vectors are modeled directly. The spectral approach based on the i-vector model [6] has proven to consistently outperform the token-based one [7]. In recent years, the viability of deploying an end-to-end neural network approach [8] to SLR has been investigated, but this frame-based technique has not outperformed the i-vector-based solution in terms of generalization performance. However, the direct connection to language cues, available in the phonotactic systems, is lost when spectral feature streams are modeled directly. In addition, short-term spectra are negatively affected by other factors, such as additive noise or the transmission channel.

Bottleneck features (BNFs) [9], [10] aim to bridge the gap between phonotactic and spectral approaches while exploiting their properties. BNFs are a feature stream generated from the linear bottleneck layer in a deep neural architecture. The neural architecture is commonly trained to recognize phonetic based classes, namely senones (tied tri-phone states) [11], from a stream of spectral features [9]. Furthermore, the neural architecture is usually fed using a long window of speech frames often spanning ten or more frames, so that the extracted BNF vector per time-step can capture acoustic relevant context, and phonetic information at the same time. The latter is related to the senone targets employed during the training phase. BNFs can then be fed into any language classifier that has already proven useful for spectral approaches. In [12], the authors observed that a bottleneck layer could preserve more phonetic information if placed closer to the output layer. That in turn has a beneficial effect on the overall SLR system. We argue the direct use of the senone-based output layer as the BNF vector could lead to top performance. Nevertheless, there are two key issues to address before employing the output layer as the BNFs, namely: (i) the BNF vectors associated with the output layer would have a very high dimension (about 3k to 9k tri-phone target labels), when a neural architecture is trained with the senone classes as targets, and (ii) the BNF vector would be intrinsically language-dependent. The latter issue could be overcome by training BNF neural architectures

Ivan Kukanov is with the School of Computing, University of Eastern Finland, Joensuu, Finland e-mail: ivan@kukanov.com.

Trung Ngo Trong is the School of Computing, University of Eastern Finland, Joensuu, Finland e-mail: trungnt13@gmail.com.

Ville Hautamäki is the School of Computing, University of Eastern Finland, Joensuu, Finland e-mail: villeh@cs.uef.fi.

S. M. Siniscalchi is with the Department of Computer Engineering, Kore University of Enna, Enna, Italy, and with the Department of Electrical S.M. Siniscalchi is also affiliated with the Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 e-mail: marco.siniscalchi@unikore.it.

V. M. Salerno is with the Department of Computer Engineering, Kore University of Enna, Enna, Italy e-mail: valerio.salerno@unikore.it

K.A. Lee is with the Biometrics Research Laboratories, NEC Corporation, Japan e-mail:kalee@ieee.org.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2964953, IEEE/ACM Transactions on Audio, Speech, and Language Processing



for multiple languages by employing stacked BNFs [10], for instance. It should be pointed out, however, that experimental evidence was reported only for two languages [10]; therefore, the viability of that approach with multiple languages has not been investigated. Furthermore, the first issue would, however, remain unsolved.

## II. Motivation

In [13], a universal acoustic characterization approach to SLR was proposed. The key idea was to describe any spoken language with a common set of fundamental units that are defined "universally" across all spoken languages. Phonetic features, referred to as speech attributes in that work, such as the manner and the place of articulation, were chosen to form the unit inventory and used to build a set of language-universal attribute models with data-driven modeling techniques. The data-driven models were used to transcribe a spoken utterance into a sequence of attributes independently of its language. Experimental evidence not only demonstrated the feasibility of the proposed techniques, but it also proved that manner and place of articulation can be used as language-independent units. It should be pointed out that several speech scientists have advocated the beneficial properties of speech attributes (phonetic features) in speech applications. For example, [14] proposed an extended front-end by appending some phonetic features to the cepstral vector, and it was shown that inter-speaker variability was reduced. In [15], a set of ANNs is used to score articulatorily-motivated features for manner and place of articulation demonstrating improved robustness against noise at low signal-to-noise ratio. In [16], a stream architecture was described to augment acoustic models based on context-dependent sub-words with articulatorily-motivated acoustic models. This work showed that articulatory features improve recognition of hyper-articulated speech.

A critical yet fundamental element of the above mentioned approaches is to build a set of data-driven models to reliably detect a collection of speech attribute cues. In fact, there are two practical configurations to deploy that set of models: (i) a set of independent 2-class classifiers can be built to detect each speech attribute of interest, and (ii) a single multi-output classifier can be implemented, simultaneously detecting all speech events. In this work, we focus on the latter configuration, because it has also the advantage of enhancing detection performance for speech attributes with insufficient training samples, as discussed in [17]. Specifically, the authors in [17] designed a single deep neural network (DNN) with multiple independent logistic regression classifiers, where those classifiers were trained independently but shared a common set of hidden layers. DNN parameters were estimated by minimizing the negative log-likelihood. In [18], a similar neural architecture was explored for phonetic feature detection, and asynchronicity among speech attributes was exploited by allowing more than one feature to be on at the same time. The mean squared error between the network output and the target output was adopted as an objective function. Those two architectural configurations actually meet the requirements of the detection framework, since an $m$-from-$N$ task is accomplished during run-time, and individual outputs can take continuous values between 0 and 1. Both studies were not concerned with the role of the objective function when attribute detection scores are used in a post-processing stage, such as lattice rescoring [19], or accent recognition [20], since the key goal was to demonstrate reliable phonetic feature detection or classification. However, a better solution, in terms of overall accuracy, could be attained by leveraging upon an objective function that may better capture the characteristics of the problem at hand, e.g., [21], [22]

Leveraging the latter intuition, we propose to cast the task of extracting speech attributes from the speech signal into a *multi-label classification* problem [24], [22]. According to the *multi-label learning* theory [25], each observation can be associated with multiple labels at the same time. Figures 1 and 2 explain the asynchronous nature of manner and place of articulation events, which are the speech attributes of interest in this work. In order to validate the viability of our solution, and provide a comprehensive set of comparing and contrasting experiments, we have tested two multi-label learning solutions. In the first solution, we model all speech attributes using a single DNN, where each output node has a sigmoid activation function. Each output node is associated with a single attribute class and produces a confidence score independently of the other output neurons. The *binary cross-entropy* (BCE) loss function is calculated for every detector in a *binary classification manner* to learn DNN parameters and the empirical expected loss is minimized. We refer to this system as the *baseline* approach. The major limitation of this solution is that the DNN emits *independent streams* of sigmoid scores in the range of $(0, 1)$ for each speech attribute. This problem was studied in the *discriminative learning* approaches for single-label classifiers [26]. The discriminative learning approach outputs the relative scores measuring the distance between a target and a competing anti-target scores (a.k.a. misclassification measure), similar to log-likelihood ratio approach in the Bayes decision theory [27]. It was shown that discriminative learning outperforms a *binary classification manner* in automatic speech recognition and applied in minimum error classification [28] and minimum verification error [29]. The second approach explores the *maximal figure-of-merit* (MFoM) [30], [31] learning solution, which allows us to approximate the metrics of interest, namely the micro-F1 and equal error rate (EER), with a differentiable function, so that gradient-based optimization algorithms can be applied to learn DNN parameters. Specifically, MFoM tries to improve the *decision boundary* [30] using the output sigmoid scores without the need of any intermediate calibration.

In this work, we combine, organize, and extend our previous findings, scattered among several research papers, and extend them in different ways putting forth a novel solution to address the SLR problem. The contributions of the present work are therefore as follows:

- We show that a low-dimensional feature vector can be deployed by leveraging universal units, such as manner and place of articulation as target classes within a DNN framework, with beneficial effects to SLR.
- Correlations among speech attributes and corresponding detectors can be captured by avoiding independent training of individual detectors. In particular, we adopt a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2964953, IEEE/ACM Transactions on Audio, Speech, and Language Processing



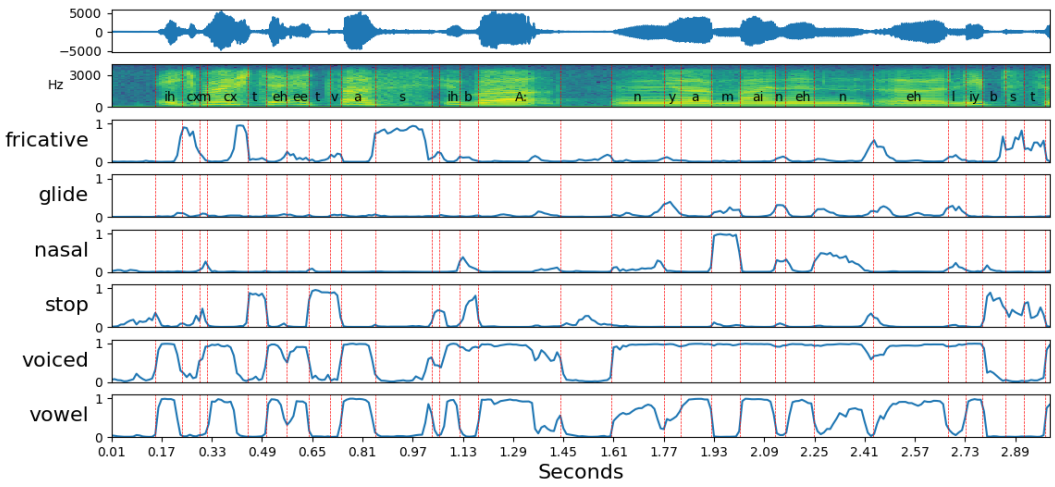


Fig. 1. Overlapping nature of the speech attributes. Human articulatory organs generate multiple events (speech attributes) in speech production. On the top, the signal and spectrogram of the phrase *"Ich möchte etwas über meinen liebst(en)..."* is shown. Under spectrogram, we depict separately several speech attributes (e.g., fricative, glide, nasal, stop, voiced, vowel), where detector tracks are produced by DNN with sigmoid output unit per speech attribute.

MFoM [30] optimization approach with a *units-vs-zeros* misclassification measure to force a single neural network to simultaneously produce detection scores for all manner and place of articulation events. We had already noticed in [32] that detectors trained in such a way turned out to be more accurate than using a separate network for manner and place. However, in [32], we trained DNN and 1D-CNN with MSE and fine-tune with MFoM-micro-F1 embedded metric. We now think of attribute detection as a single multi-label task, and we proposed *units-vs-zeros* misclassification measure special case for multi-label classification within the MFoM mathematical framework. In particular, we improve the MFoM framework by training the deep model from *scratch* without initial pre-training, what was instead done in [32].

- In [33] and [34], it was proven that state-of-the-art results can be delivered through MFoM and recurrent neural networks for a multi-label audio tagging task. This paper explores a modified version of the convolutional recurrent neural network (CRNN) [34] with *time distributed* output layer and MFoM training [34] for detecting attributes in SLR applications. Section V gives more details.
- We demonstrate that improvements at a speech attribute level positively affect the SLR performance with a series of experiments on the NIST LRE 2017 task.

## III. Speech Attribute Modeling

### A. Speech Attributes

The problem of *attribute detection* is formally described in the *automatic speech attribute transcription* (ASAT) framework [35], [36]. ASAT is a bottom-up detection-based framework, where speech attributes are extracted using data-driven modeling techniques without physical real-time magnetic resonance imaging methods (rtMRI) [37]. The main goal of the project was to promote the development of new approaches based on the detection of speech attributes and phonological knowledge integration. Several successful applications of the framework have been proposed in different domains of speech processing, such as phoneme recognition [38], foreign accent recognition [39], language recognition [2]. Speech attributes of interest are mainly *manner of articulation*, namely **fricative, glide, nasal, stop, vowel, voiced**, and *place of articulation*, namely **coronal, dental, glottal, high, labial, low, middle, palatal** and **velar**. In the present work, we decided to add the **voiced** class to the manner of articulation. Whereas the voiced class is separated from the manner and the place of articulation according to the *voice-place-manner* (VPM) [40] model.

Speech attributes can be obtained for a particular language and shared across many different languages, and those attributes can thereby be used to derive a *universal set of speech units* [41], see Fig. 1 with detected speech attributes and relation to phonemes. We can observe that one phoneme can belong to several attribute classes; therefore, a stream of attribute labels can be assigned to a single phoneme observation according to phonetic knowledge [42]. Phonemes possess several physiological articulation features, since movements of several vocal organs are usually required, and sound rises in different parts of a vocal tract. For instance, phoneme /ih/ is detected as *voiced*, *vowel* (at 0.16 sec) and phoneme /m/ as *nasal*, *voiced* (at 1.93 sec).

Fig. 2 shows the connection between different speech attribute classes. It should be noticed that pair *voiced-vowel* is the most frequent in the OGI-TS [23] database (more than 100k pairs of observations). Moreover, the *voiced* class is paired with almost all attributes. On the other side, the *glottal* attribute has the lowest number of combinations with other classes. The fact that some articulatory classes appear with other classes led us to consider the multi-label classification as the problem formulation in our case.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2964953, IEEE/ACM Transactions on Audio, Speech, and Language Processing



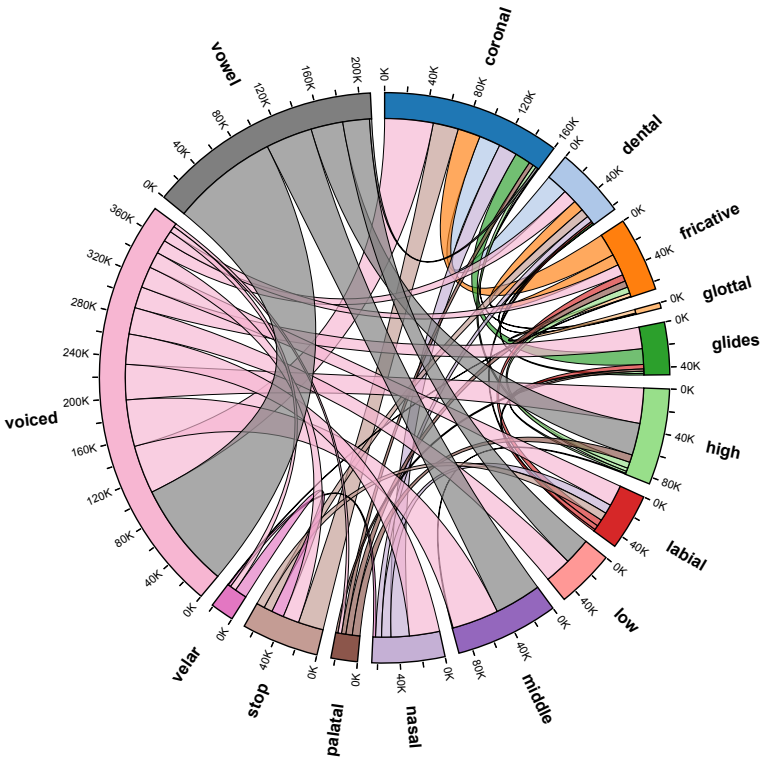


Fig. 2. Chord diagram shows the interconnection of attribute classes. On the radius, it is shown the number of particular attribute observations, those numbers were crafted from the OGI Multi-language Telephone Speech (OGI-TS) corpus [23]. The thickness of the connecting branch between a pair of attribute classes shows how many times the pair occurs in the OGI-TS speech corpus.

### B. Multi-label Classification Settings

As mentioned above, one phoneme can be mapped into several articulatory attributes [42], and we can treat the attribute detection problem as a *multi-label* classification task. Articulatory attributes have diverse acoustic nature: some attributes are impulsive and have a low frequency (e.g., *stop* attribute); whereas, others have broadband frequency characteristics (e.g., *voiced*). Therefore, an automatic system should extract features that benefit both of these properties. Conventional parameterization of raw audio input signals is in the form of matrices comprising of consecutive frames (log-Mel filter banks) [43]. We denote the matrix of observations as $\mathbf{X} \in \mathbb{R}^D$ of size $D = [D_{\text{FB}} \times D_{\text{T}}]$, where $D_{\text{FB}}$ is the number of filter banks and $D_{\text{T}}$ is the number of consecutive frames taken from a speech utterance. Each observation matrix $\mathbf{X}$ of speech frames is associated with a corresponding binary vector $\mathbf{y} \in \{0,1\}^M$, which has several unit marks corresponding to attribute class labels, e.g., $\mathbf{y} = (1, 0, \ldots, 1, 0)^\top$. In this work, two types of speech attributes are modeled, namely manner (6 classes, $M = 6$) and place (9 classes, $M = 9$) [42]. The training set of labeled speech utterances is defined as $\mathbb{T} = \left\{ (\mathbf{X}_i, \mathbf{y}_i) | \, i = \overline{1, N} \right\}$.

In the training phase, the temporal context of filter bank features $\mathbf{X}_i$ are fed to the artificial neural network, see Fig. 3. The number of output units is equal to the number of attribute classes (6 or 9).

## IV. Multi-label Classification

The *binary cross-entropy* (BCE) loss function is commonly used for optimizing neural network parameters in multi-label acoustic events detection [44]. BCE is defined as follows,

$$J_{\text{BCE}}\left(\mathbb{W}|\mathbb{T}\right) = \frac{1}{N}\sum_{i=1}^{N}\big\{-\mathbf{y}_i^\top \log\left(\mathbf{g}_i\right) - \\ - \left(1-\mathbf{y}_i\right)^\top \log\left(1-\mathbf{g}_i\right)\big\}, \qquad (1)$$

where the network parameters are $\mathbb{W} = \left\{\mathbf{W}_n | n = \overline{0, L}\right\}$, with $L+1$ layers; $\mathbf{g}_i \in \mathbb{R}^M$ is the vector of output scores corresponding to input features $\mathbf{X}_i$. The $k$-th element of the vector $\mathbf{g}_i$ is the output of $k$-th unit of network

$$g_k\left(\mathbf{X}_i; \mathbb{W}\right), \quad k = \overline{1, M}, \qquad (2)$$

where $g_k$ is known as *discriminant function* [45] for the class $C_k$. In multi-label classification, thresholding is applied to the neural network output as a decision rule for binarization to choose several class candidates for the current input observation. In the baseline DNN system, we use the sigmoid output scores as *discriminant functions* for a class $C_k,\ k = 1, \ldots, M$.

### A. Limitations of the BCE

In multi-label classification, the outputs of the classifiers are typically modeled independently, i.e., the detection problem for each class is considered as an independent binary cross-entropy task. The global error is then obtained as the sum of the binary predicted probability for each label and averaged

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2964953, IEEE/ACM Transactions on Audio, Speech, and Language Processing



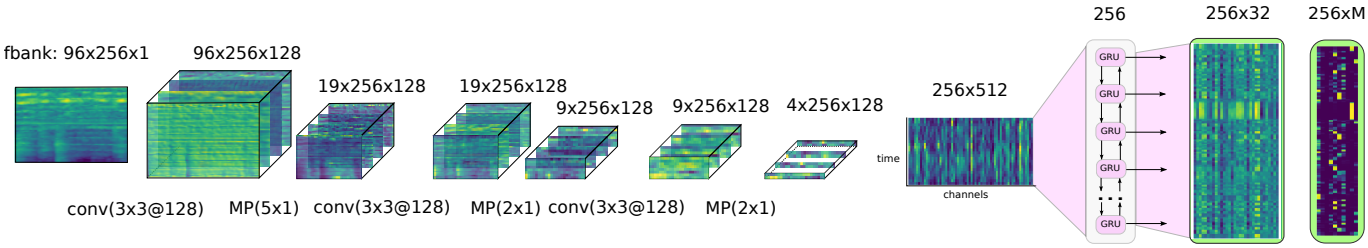


Fig. 3. Multi-label architecture using convolutional recurrent neural network (CRNN). Sequence of convolutions and max-pooling is followed by bi-directional gated recurrent unit (Bi-GRU), which is unfolded on the figure. The output decision layer has dimension of $256 \times M$, where $M$ is the number of speech attribute classes ($M = 6$ for manner, $M = 9$ for place or $M = 15$ for fusion of articulatory attributes). We optimize either binary cross-entropy or MFoM-micro-F1, MFoM-EER objective functions for the same network architecture.

across the number of available samples $N$, and the number of labels $M$. By optimizing the BCE error criteria, "the distance between what the network believes the distribution should be, and what the teacher gives as target" is minimized, i.e., the Jensen-Shannon divergence is minimized [46]. Considering auxiliary information, such as the interconnection among labels, helps to improve the classification accuracy of the multi-label classification model, e.g. [47]. In addition, the key limitation of the BCE loss is that it does not allow the inclusion of task specific performance metrics during to be optimized directly.

### B. Objective functions based on performance-metrics

In [48], optimization of the *infomax criterion* [49] and its relation to *balanced error rate* (BER) [50], F1 and cost sensitive objectives is studied. Universal lower and upper bounds, namely Fanos and Hellmans bounds [48], are obtained for BER, F-score and cost-sensitive risk. The main outcome of the study was that conditional entropy minimization does not guarantee neither the minimization of the cost sensitive risk, nor the maximization of the F-score. The cost of the errors on different samples is different when dealing with skewed datasets, i.e., imbalanced datasets, and thereby cost-sensitive risk, or F-score are more suitable in those scenarios [51]. In [48], numerical examples confirming that the minimization of the conditional entropy is inconsistent with the cost-sensitive risk, and the F-score were given. Moreover, conditional entropy minimization may even lead to contradictory results: Reducing the entropy degrades the F-score. The latter implies that conditional entropy optimization may even lead to a poor data-driven modeling process when F-score, or cost-sensitive performance measures are used. The question concerned with finding a consistent information measure for F-score is still open [48] and is related to the *non-decomposable objective functions* problem. The interested reader is referred to Appendix A for more details on non-decomposable objective functions.

The beneficial effects of adopting performance-metrics objective function is also demonstrated by recent studies. Fore example, the optimization of the area under the ROC curve, $F_\beta$, precision at fixed recall, or mean average precision were investigated for deploying a ranking-based system in [52]. The approach was applied to large-scale image classification tasks, such as ImageNet [53], and it was demonstrated that models trained leveraging non-decomposable objective functions can outperform corresponding models built with conventional decomposable objective functions, such as cross-entropy. In [54], better speaker verification systems could be deployed by adopting a performance-based objective function, such as DCF, AUC, EER. More in detail, the authors proposed an end to end objective function based on DCF performance in combination with FPR and FNR, which allowed to train a score decision threshold directly during backpropagation. The latter is indeed a promising direction for self-calibrated approaches. [33] demonstrated that a *units-vs-zeros* misclassification measure can improve discrimination in multi-label acoustic events detection task.

On the one hand, objective functions based on performance metrics are difficult to optimise, as discussed in [55], [54]. On the other hand, those objective functions allow to incorporate task specific performance metrics in the backpropagation optimization process. Therefore, we no longer rely on indirect error rate optimization in the hope that cost-sensitive performance is improved as well. Finally, auto-calibration training methods could be derived in the future based on non-decomposable objective functions. In the next section, we describe in detail the MFoM framework that allow us to take into account the performance metric used for assessing the task at hand. The experimental evidence reported in Section VII demonstrate the effectiveness of our idea.

## V. Multi-label recognition with MFoM

In this section, we present the key ingredients to deploy a differentiable objective function based on micro-F1 and EER within the MFoM framework, namely: *discriminant functions*, *misclassification distance measure* and *smooth error count*.

### A. Discriminant Function

The choice of a proper discriminant function (2) depends on the nature of the classifier, and the task at hand. Discriminant functions are defined on the classifier parameters set $\mathbb{W}$. The goal is to find the optimal set of parameters that minimizes the objective function (e.g., binary cross-entropy in (1)), and the discriminant functions must satisfy the decision rule for any sample $\mathbf{X}_i$ of class $C_k$ as follows

$$g_k\left(\mathbf{X}_i; \mathbb{W}\right) > g_j\left(\mathbf{X}_i; \mathbb{W}\right), \tag{3}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2964953, IEEE/ACM Transactions on Audio, Speech, and Language Processing



where $k \in \mathbf{y}_{\{1\}}$ is the set of indices corresponding to **1** in the label vector, $\mathbf{y}$; accordingly $j \in \mathbf{y}_{\{0\}}$ is the set of indices corresponding to **0** in $\mathbf{y}$. The condition in (3) has a unique $k$ for any sample $\mathbf{X}_i$ in case of *single-label* classification, because $\mathbf{X}_i$ belongs to a single class $C_k$; whereas, $k$ is a set of several indices for any particular $\mathbf{X}_i$ for *multi-label* classification.

### B. Misclassification Measure

The idea behind a *misclassification measure* is to represent a decision rule (3) in a functional form, which is suitable for a gradient based optimization, see Fig. 4. Those decision rules provides the classifier with an additional information about the relationships among classes. Different families of misclassification measures for the single-label classification case are described in [26], [28]. Our contribution to the misclassification measures for *multi-label* classification was presented in [32], [34], and we here focus on the *units-vs-zeros* misclassification measure, $\psi_k$, from [32] that measures the misclassification for the current class, $C_k$, as follows:

$$\psi_k = -g_k + \frac{1}{\eta} \ln \left( \frac{1}{|\mathbf{I}|} \sum_{j \in \mathbf{I}} e^{\eta g_j} \right), \tag{4}$$

$$\begin{cases} \text{if } C_k \text{ is } 1 \Rightarrow \mathbf{I} = \mathbf{y}_{\{0\}}, \\ \text{if } C_k \text{ is } 0 \Rightarrow \mathbf{I} = \mathbf{y}_{\{1\}}, \end{cases} \tag{5}$$

where $\psi_k$ is defined 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 $\mathbf{y}$; the discriminant functions are indicated by $g_k$, and $g_j$. Finally, $\eta$ is a positive real-valued smoothing constant.

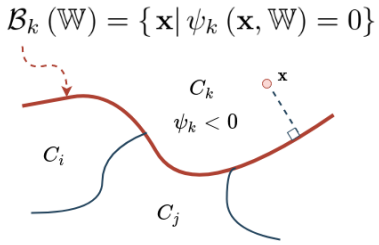


Fig. 4. Graphical interpretation of the misclassification measure. If misclassification measure $\psi_k = 0$ for a sample $\mathbf{x}$, then this sample is on the decision boundary $\mathcal{B}_k$. Otherwise, the absolute value of the misclassification measure defines a distance to the decision boundary and the sign tells the decision: $\psi_k < 0$ means a sample belongs to the class $C_k$, else it is misclassified.

On the right-hand-side of (4), the first term is referred to as the target model, and the second term is the geometrical mean (a.k.a. Kolmogorov mean [56]) of the competing models. Varying the parameter $\eta$ enables the emulation of various decision rules. In the extreme case, when $\eta \to +\infty$, the geometrical average becomes a maximum metric [56], i.e., it converges to the highest score among all competing classes. The conditions in (5) describe an explicit incorporation of the label information into the *units-vs-zeros* measure (4). For the current class, $C_k$, labeled as 1, the competing models, $C_j$, are only those indicated with the label 0, and vice versa, if $C_k$ is labeled as 0. Therefore, (5) properly formulates the decision inequalities (3) when a sample $\mathbf{X}$ belongs to several classes at the same time.

The sign of the misclassification measure indicates the correctness of classification: $\psi_k(\cdot) < 0$ means that the predicted class is correct; whereas, $\psi_k(\cdot) > 0$ implies an incorrect decision. The absolute value of the $\psi_k$ quantifies the margin between current sample $\mathbf{X}$ and the decision boundary (see Fig. 4). The $\psi_k(\cdot) = 0$ defines the decision boundary between the class $C_k$ and the rest. In the training phase, $\psi_k(\cdot)$ is adjusted to make a right decision for the samples which are on the boundary $\mathcal{B}_k$ (i.e., $\psi_k(\mathbf{X}) = 0$) or misclassified samples (i.e., $\psi_k(\mathbf{X}) > 0$).

### C. Smooth Error Count

The third component of the MFoM framework is the *smooth error count*, which is needed for the approximation of discrete performance measures based on discrete error counts (i.e., false positive and false negative statistics). We therefore introduce a smooth (differentiable), and monotonic approximation function that squeezes the output of the *misclassification measure* to the $[0, 1]$ range. That squeezing function can be a sigmoid, a hinge, an exponential, or any other smooth function. In this paper, the sigmoid function is selected to approximate the discrete error count of the misclassified samples; it is a smoothed version of the error step function [57], applied to the measure (4):

$$l_k = \frac{1}{1 + \exp\left[-\alpha_k \psi_k - \beta_k\right]}, \tag{6}$$

where $k = \overline{1, M}$ is the class index, and $\alpha_k$ and $\beta_k$ are real valued parameters of the scale and shift transformation, respectively. For the analysis of the $\alpha_k$ and $\beta_k$ parameters, an empirical method presented in [30] is used to find them. From a deep learning point of view, we can interpret the linear transformation ($\alpha_k$ and $\beta_k$) of the *misclassification measure* as an additional layer of a network. Hence we propose the optimization of those parameters in a way similar to the batch normalization technique in [58], when the error of the objective function, $E$ is backpropagated through $\alpha_k$ and $\beta_k$ as well:

$$\frac{\partial E}{\partial \alpha_k} = -\frac{\partial E}{\partial l_k} \cdot \psi_k, \tag{7}$$

$$\frac{\partial E}{\partial \beta_k} = -\frac{\partial E}{\partial l_k}. \tag{8}$$

It is worth to remark that in the binary cross-entropy (1), the objective of learning is to minimize the number of errors by reducing the entropy, and neural network scores $\mathbf{g}$ do not posses the class interconnection information. Whereas, the smooth error count (6) encapsulates the misclassification measure (4) with the implicit class relationships, and that forces a neural network to learn task specific information. Moreover, the smooth error count will be optimized by the proposed performance objective in the next Section.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2964953, IEEE/ACM Transactions on Audio, Speech, and Language Processing



### D. Approximation of Micro-F1 Objective

One of the most common performance metric for multi-label classification is the micro-$\mathrm{F}_1$ (or micro-averaged F1) [59], [60], which is the harmonic mean of precision, P, and recall, R, and can be expressed as a function of the discrete count of true positives, $\mathrm{TP}_\mathrm{k}$, false positives, $\mathrm{FP}_\mathrm{k}$, and false negatives, $\mathrm{FN}_\mathrm{k}$, [59] as follows:

$$\mathrm{F}_1 = \frac{2 \cdot \mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}} = \frac{\sum\limits_{\mathrm{k}=1}^{\mathrm{M}} \mathrm{TP}_\mathrm{k}}{\sum\limits_{\mathrm{k}=1}^{\mathrm{M}} \left(\mathrm{TP}_\mathrm{k} + 2 \cdot \mathrm{FP}_\mathrm{k} + \mathrm{FN}_\mathrm{k}\right)}. \tag{9}$$

As discussed above, the key ingredients of the proposed MFoM framework are: a) the discriminant functions, $\mathbf{g}_k$ in (2), which are the sigmoid activations in the last layer of the neural architecture, b) a misclassification measure (4), and c) smoothed error count (6). With those three elements, we can now express the micro-F1 function in terms of those three entities within the deep neural network paradigm. We introduce a smooth approximation of the error counts of true positive, false positive, and false negative outcomes in (9) following [30]:

$$\mathrm{TP}_\mathrm{k} \approx \sum_{\mathbf{x} \in \mathbb{T}} \left(1 - \mathrm{l}_\mathrm{k}\right) \cdot \mathbf{1}\left(\mathbf{x} \in \mathrm{C}_\mathrm{k}\right), \tag{10}$$

$$\mathrm{FP}_\mathrm{k} \approx \sum_{\mathbf{x} \in \mathbb{T}} \left(1 - \mathrm{l}_\mathrm{k}\right) \cdot \mathbf{1}\left(\mathbf{x} \notin \mathrm{C}_\mathrm{k}\right), \tag{11}$$

$$\mathrm{FN}_\mathrm{k} \approx \sum_{\mathbf{x} \in \mathbb{T}} \mathrm{l}_\mathrm{k} \cdot \mathbf{1}\left(\mathbf{x} \in \mathrm{C}_\mathrm{k}\right), \tag{12}$$

where $\mathbf{1}(\cdot)$ is the indicator function of the logical expression, $\mathbf{x}$ is a training sample from a dataset $\mathbb{T}$. Thus, a differentiable micro-F1 is eventually obtained

$$E_\mathrm{F}(\mathbb{W}) = 1 - \mathrm{F}_1(\mathbb{W}), \tag{13}$$

where $\mathbb{W}$ is a network parameters. Furthermore, we minimize this objective function during a neural network training phase. For Jacobian inference and analysis of the objective function (13), see Appendix A-A.

### E. Approximation of EER Objective

In this section we infer a smooth approximation of the discrete EER within the MFoM framework. The EER is expressed through two types of errors, namely a *false negative rate* (FNR) and a *false positive rate* (FPR). $\mathrm{FNR}(\mathrm{t})$ and $\mathrm{FPR}(\mathrm{t})$ are increasing and decreasing functions of a threshold $t \in [0, 1]$, and the value of EER is defined on those intersection. The lower the value of the EER is, the better the performance of a system is. The EER is defined, $\triangleq$, as follows:

$$\mathrm{EER}(\mathrm{t}^*) \triangleq \mathrm{FNR}(\mathrm{t}^*) = \mathrm{FPR}(\mathrm{t}^*), \tag{14}$$

with the optimal threshold $t^*$, where

$$\mathrm{FNR}(\mathrm{t}) = \frac{\mathrm{FN}(\mathrm{t})}{\mathrm{P}}, \quad \mathrm{FPR} = \frac{\mathrm{FP}(\mathrm{t})}{\mathrm{N}}, \tag{15}$$

and P, and N are the total numbers of positive and negative samples, respectively. The optimal threshold for the EER is $t^* \in [0, 1]$. The criterion for the optimal threshold is defined through the following *intersection condition*

$$t^* = \underset{t}{\operatorname{argmin}} \left|\mathrm{FNR}(\mathrm{t}) - \mathrm{FPR}(\mathrm{t})\right|. \tag{16}$$

The goal is to develop an objective function that directly optimizes the EER. The EER can be parametrized with a neural weights, $\mathbb{W}$, and represented as an optimization problem. Wit the equality (14) as the *intersection condition*, we have two natural alternatives for EER optimization, namely

$$\begin{aligned} &\mathrm{FPR}(\mathbb{W}) \to \min_{\mathbb{W}}, \quad \text{or} \quad \mathrm{FNR}(\mathbb{W}) \to \min_{\mathbb{W}}, \\ &\text{subject to} \quad \left|\mathrm{FNR}(\mathbb{W}) - \mathrm{FPR}(\mathbb{W}) = 0. \right. \end{aligned} \tag{17}$$

The problem (17) is a conditional optimization, and we can reformulate it as a **Lagrangian dual problem**. Therefore, we obtain the EER as the objective function with model parameters $\mathbb{W}$ as follows

$$\begin{aligned} E_\mathrm{EER}\left(\mathbb{W}\right) = \mathrm{FPR}\left(\mathbb{W}\right) + \\ + \lambda \left|\mathrm{FNR}\left(\mathbb{W}\right) - \mathrm{FPR}\left(\mathbb{W}\right)\right|, \end{aligned} \tag{18}$$

where FPR, and FNR are smoothed false positive, and false negative rates, respectively, and $\lambda \geq 0$ is Lagrange multiplier, a.k.a. *dual variable*. As the concept testing, we set $\lambda = 1$, and the cost of the minimization of FPR and the intersection condition (FNR and FPR) are equivalent in (18). In this formulation, the intersection condition is a regularization condition for FPR minimization. Discrete FPR, and FNR are approximated using smooth false positive (11), and false negative (12) counts, as follows

$$\mathrm{FPR}_\mathrm{k} = \frac{\mathrm{FP}_\mathrm{k}}{\mathrm{N}_\mathrm{k}}, \tag{19}$$

and

$$\mathrm{FNR}_\mathrm{k} = \frac{\mathrm{FN}_\mathrm{k}}{\mathrm{P}_\mathrm{k}}, \tag{20}$$

in order to simplify the notation, we omit parameter $\mathbb{W}$. Finally, the MFoM-EER objective function for each class $k = \overline{1, M}$

$$E_k = \mathrm{FPR}_\mathrm{k} + \lambda \left|\mathrm{FNR}_\mathrm{k} - \mathrm{FPR}_\mathrm{k}\right|, \tag{21}$$

and the averaged *class-based* MFoM-EER is minimized

$$E_\mathrm{EER} = \frac{1}{M} \sum_{k=1}^{M} E_k. \tag{22}$$

### F. Proposed MFoM-based Neural Architecture

MFoM-based objective functions are MFoM-micro-F1 and MFoM-EER, i.e., objective functions with embedded performance measures (F1 and EER, respectively) that are optimized leveraging the back-propagation algorithm. In order to isolate the effect of the MFoM-based learning, we train the same neural architecture shown in Fig. 3 using either BCE, or MFoM. Differences between the two neural models can therefore be directly associated with changes in the objective functions, learning rate, gradient optimization techniques, and network output activation functions. The CRNN model to be

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2964953, IEEE/ACM Transactions on Audio, Speech, and Language Processing



optimized with MFoM-based objective function can have randomly (*glorot-uniform* [61]) initialized weights. In this case, MFoM is applied from scratch. We could also start MFoM training using a seed CRNN learned using BCE algorithm, and we could think of such an approach as a parameter fine-tuning. As shown in [32], fine-tuning with MFoM improves the baseline model performance. In this work, we managed to attain the same performance using MFoM from *scratch*, which obviously reduces the training effort.

The MFoM pipeline calculation (see Appendix A, Fig. 8), for the forward pass of the backpropagation is based on the network output scores $\mathbf{g}$ from (2), then the misclassification measure (4) and smooth error count function (6) are obtained. The MFoM, micro-F1 from (13) or EER from (22), depends on the intermediate statistics, i.e., approximated smoothed counts TP, FP and FN from (10) - (12). Those statistics are accumulated over every mini-bath $\mathbb{T}$ for each time frame (40ms). Next, either *micro-averaging* (instance-based) or *macro-averaging* (class-based) averaging strategy [62] is applied.

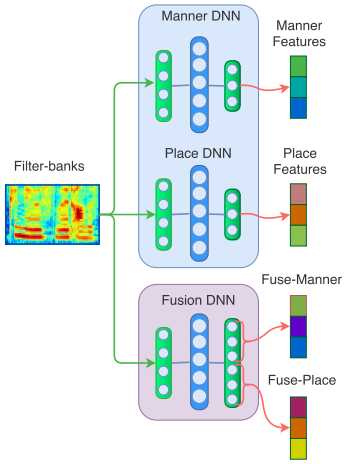


Fig. 5. Four types of speech attribute features. We train three separate neural networks: *manner* and *place* DNNs, and fusion for joint training. In the fusion DNN model, some of the output units are in charge of detecting manner attributes (*Fuse-Manner*) while the others are responsible for detecting place attributes (*Fuse-Place*) [1].

## VI. Experimental Setup

### A. Speech Attribute Classifier Training

*1) Groundtruth for Multi-label Speech Attributes:* Speech attribute models (see Fig. 5) are trained on the *stories* subset of the OGI Multi-language Telephone Speech (OGI-TS) corpus [23]. This dataset has audio recordings for six languages: *English*, *German*, *Hindi*, *Japanese*, *Mandarin*, and *Spanish*. Time-aligned phonetic labels are provided for those recordings. In order to train universal and robust articulatory attributes across languages, we pool all recordings for six languages to get $5.57$ hours of training and $0.52$ hours of test data. OGI-TS dataset has the time-aligned phoneme labels, but a ground-truth information is needed in order to train attribute detectors. We convert phoneme labels into corresponding attribute classes according to the phonological tables in [42]. In this work, we consider attribute detection as a multi-label classification problem, that is, our task requires to find both onset and offset time for multiple overlapping attribute classes in the input recording.

Following [33], convolutional recurrent neural networks (CRNNs) are used as building blocks of our multi-label classification system, see Fig. 3. However, we preserve here the time dimension of the input Mel-filter bank feature through all network layers in order to align input features with target labels at each time frame. We compare two different schemes to train our multi-label attribute classifiers (see Fig. 5): (i) Two independent neural architectures, one for *manner*, and one for *place* versus (ii) *a single fusion neural architecture* to model simultaneously manner and place attributes. The last layer of the fusion network emits joint scores for manner, and place attributes. Therefore, four types of features can be evaluated: (i) *manner*, (ii) *place*, (iii) *fuse-manner*, and (iv) *fuse-place*.

*2) BCE-based Neural Architecture - Baseline:* The input to the CRNN in Fig. 3 is a feature matrix of $\mathbf{X} \in \mathbb{R}^{D \times T}$, where $D = 96$ is the dimension of log-Mel filter banks spanning from 0 to 4 kHz Nyquist frequency (sampling rate is at 8 kHz), and the context window spans $T = 256$ time frames. In [63], it is reported that a wider context window is beneficial for polyphonic sound event detection in real-life environments. Indeed, a wider context allows effective modeling of longer sound events, and events correlations, which in turn leads to a better modeling of the temporal information.

In the CRNN, a 2-dimensional convolutional layer is trained directly on raw log-Mel filter bank features $\mathbf{X}$, and every convolutional output is passed through an exponential linear unit (ELU) [64] activation function. Three convolution transformations with $(3 \times 3)$ filters followed by a max-pooling operation with $(5 \times 1) \rightarrow (2 \times 1) \rightarrow (2 \times 1)$ kernels are used in our CRNN. Nevertheless, max-pooling is carried out on the frequency axis only in order to preserve the time information for final attribute detection. In fact, the time dimension $T$ remains unaltered through the whole network, and that preserves the alignment between input frames $\mathbf{X}$, and target labels $\mathbf{y}$. Next, the processed input features are sent to *bi-directional gated recurrent units* (Bi-GRUs) based block. In our architecture, the convolution layers extract relevant local features and smooth audio distortions out; whereas, the Bi-GRUs block models the temporal context information. In other words, the convolutional layers reduce the effect of time-frequency distortions and extract stable and denoised features, but those features lack of a longer temporal context summarization effect. The recurrent part is therefore used to model temporal information (theoretically unlimited) not handled by the convolutional block. It is worth pointing out that the authors in [63] have shown that RNNs suffer from frequency domain noise and pitch-shifting. The combination of both CNN and RNN architectures improves thus acoustic events detection.

The Bi-GRU block returns a sequence of hidden state vectors of 32 dimension per time frame, which is further processed by a *time distributed* fully-connected layer having a *sigmoid* output unit per each articulatory attribute class (or

[1]The project source code for training attributes can be found here www.github.com/Vanova/mfom_attribute_detection

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2964953, IEEE/ACM Transactions on Audio, Speech, and Language Processing



$\mathbf{g} \in \mathbb{R}^M$ vector of discriminant functions in (2) per time frame). The output layer has a dimension equal to $T \times M$, where $M$ is the number of speech attributes (6 for the manner and 9 for the place, or 15 for the fusion). The model generates confidence scores for $T$ consecutive frames at once for every input $\mathbf{X}$. The *binary cross-entropy* (BCE) objective function (1) is employed to train the neural architecture, which is referred to as the ***baseline*** system. During training, we slide the features context window with 70% overlapping across the audio file. When the end of file is reached, the next file is randomly selected up till a batch size of 32 frames is reached. At each epoch, our neural model is exposed to all available audio files. For validation and testing, overlapping is not used.

In this work, we calculate segment-based evaluation metric [62] on the test set, namely equal error rate (EER). The segment length is a single time frame (40 ms). For every consecutive time frame of an input feature matrix $\mathbf{X}$, the CRNN model produces $\mathbf{g}$ vectors of confidence scores for each class $k = \overline{1, M}$ as in (2). The performance EER is calculated for each articulatory attribute class and class-wise averaged to obtain the AvgEER. The AvgEER for the baseline is reported in the first column in Tables I, and II.

*3) MFoM-based Neural Architecture - Proposed:* The proposed neural architecture has the same architecture as the baseline model. The research interest is in the optimization capability of the MFoM objective functions. Therefore, in the baseline architecture, we make a minimal changes: instead of BCE, the MFoM-based objective functions are optimized while the sigmoid output activation function is replaced with *hyperbolic tangent*.

### B. Spoken Language Recognition System

*1) NIST LRE17 Corpus:* The availability of large corpora in speech processing has been one of the major driving forces advancing speech technologies [66]. The NIST 2017 language recognition evaluation (LRE17) dataset is the most recent effort to advance research in LRE. The challenge, as described in the evaluation plan [65], builds on the history of the LRE campaigns, and it shares many features with the previous challenges. However, there are two major differences that pose challenges to the speech community, namely:

- The inclusion of *VAST* utterances in development set and evaluation set. Those audio recordings were extracted from video data in a much different encoding and channel variations compared to traditional telephone speech available in *MLS14* corpus.
- The use of normalized cross-entropy ($C_{norm}$) as performance metrics. The evaluation process calculates $C_{norm}$ for each language under two assumed prior probabilities $P_{true} = 0.5$ and $P_{true} = 0.1$. The final score is the average of all those values.

We want to assess the ability of each technique in domain adaptation, i.e. match the performance on both *MLS14* and *VAST* utterances; therefore, our strategy is to limit the amount of VAST material during training by randomly picking only 30% of the development to form the training set. The held-out material, referred to as *validation set*, is then used for early-stopping, tuning hyper-parameters, validation, and as an alternative evaluation for the system performance. We would also like to emphasize that the evaluation set has not been touched, and it is used during scoring phase only. To sum up, there are 17425 files for training, 2440 files for validation and 25449 files for evaluation.

*2) SDC & Mel-Spectrogram Speech Features:* We use i-vector extractor [66] to build a basic spoken language recognition system. Starting with a 512-dimensional Fourier transform on 25 (ms) frames and 10 (ms) step length, we extracted two sets of acoustic features:

- 40-dimensional Mel-filter banks spectrogram (MSpec) together with its delta and delta-delta coefficients.
- shifted delta coefficients (SDC) [67] were calculated on 7 consecutive frames of 7-dimensional cepstral coefficients (MFCCs). The delta coefficients are calculated for every 3 frames, and all 49-dim delta features are concatenated with original MFCCs to form 56-dim SDC features.

We train a *universal background model* (UBM) for every type of features with 2048 Gaussians with diagonal covariances. The diagonal UBM was deployed to build the total variability matrix and extract the 400-dimensional i-vectors. *Within-class covariance normalization* (WCCN) [68] and linear discriminant analysis (LDA) are applied to project the i-vectors onto a sub-space where inter-dialect variability is maximized and intra-dialect variability is minimized.

As a language classifier, we employed the support vector machine (SVM) [67]. We train a multi-class SVM according to a one-vs-one scheme, which handles a multi-classification task while dealing with the non-linearity of speech and language representation [67]. We empirically select radial basis function (RBF) kernel after it outperformed other options including: linear, polynomial, and sigmoid kernel. *This post-processing pipeline for the features (MSpec and SDC) and classification SVM method are repeated for all experiments same backend to ensure the comparable results.*

*3) Deep Bottleneck Features Based i-Vector:* i-Vectors can be built also around bottleneck features, as discussed in the introductory section. Deep bottleneck features [9] are trained over 13-dimensional MFCC features concatenated with delta, and delta-delta coefficients. Those features are generated from the *Switchboard-1*, and *Fisher* corpora ($\approx$ 2000 hours). Those features are then processed using a per utterance mean and variance normalization and stacked with 10 past and 10 future frames to form a 21-contextual feature vector. The DNN used to extract bottleneck features has seven hidden layers with 2048 units, and a bottleneck layer with 80 units. The bottleneck layer is placed two layers before the output one. We have used ReLU activation followed by a re-normalization that scales the activations RMSE to 1.0. For the bottleneck layer, however, we have only applied re-normalization. The output layer has 8700 targets, and each target corresponds to a senone obtained with an off-the-shelf speaker-independent automatic speech recognition system. The 80-dimensional bottleneck features are employed to generate i-vectors for each spoken utterance. An energy-based voice activity detection (VAD) routine is applied to the raw bottleneck features in order to remove silence frames. Finally, those i-vectors employed in the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2964953, IEEE/ACM Transactions on Audio, Speech, and Language Processing



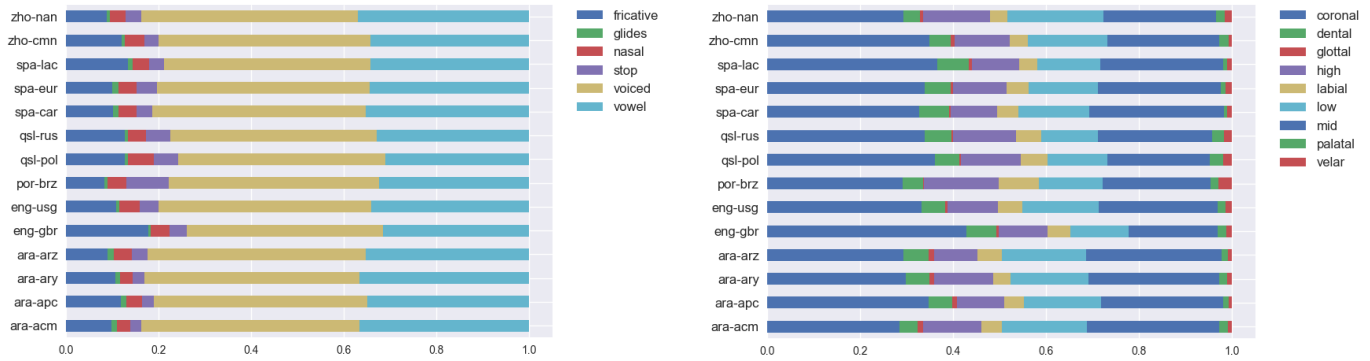


Fig. 6. The statistical mean values (i.e., every patch on the bars) of speech attribute detectors per each language are calculated on the NIST LRE17 corpus [65]. Those mean values show the difference between target languages (14 target languages) in terms of manner attributes (left figure) and place attributes (right figure). The place attributes better capture the differences across the languages and benefit the recognition.

language classifier for accomplishing the language recognition task. The architecture of the language recognition backend is the same of that used for the SDC, and MSpec solutions.

## VII. Experimental Results

### A. Attribute Detectors Analysis

Table I presents the performance of attribute detection. The first column (BCE) shows the EER values when the BCE objective function is employed. The next four columns refer to MFoM-F1 and MFoM-EER performances when the attribute detectors are trained within the MFoM framework. The last column (MFoM-F1 [32]) displays results from our previous work for comparison purposes. We refer to the performance attained by applying MFoM over a seed model built using the BCE objective function as *tuning*. When the parameters of the neural networks are randomly initialized, we refer to such a configuration as a *scratch*. For both objective functions (MFoM-EER and MFoM-F1) the training with pre-initialized weights (*tuning*) outperforms models randomly initialized (*scratch*), even though the scratch configuration is the most interesting since speeds up the deployment phase. We can also notice by inspecting Table I values that the *fusion* architecture, shown in Fig. 5, seems to give a consistent performance improvement across attributes (manner and place) and training schemes (BCE and MFoM). In particular, *fuse-manner* and *fuse-place* detectors have superior accuracy compared to the attribute detectors independently trained with stand-alone neural architectures (i.e., *place* and *manner* in Fig. 5). The current solution also outperforms the result obtained in our previous work [32], where the 1D-CNN network was trained with the mean squared error (MSE) objective and fine-tuned with the MFoM-F1. A more general performance picture can be shown by the detection error tradeoff (DET) [69],[70], i.e., curves of the false rejection rate (FRR) versus false acceptance rate (FAR), see Fig. 7. It is important for practical applications to compare a discrimination capability of the systems for different score thresholds. Fig. 7 shows the performance of the current attribute system trained with MFoM-EER (*Manner* and *Place*) and with MFoM-F1 objective (*Manner** and *Place**) from the previous work [32]. A confident improvement of the proposed system across all *operating points* can be seen.

TABLE I

*Performance of speech attribute CRNN models (manner, place and fusion). Attribute detectors' models are trained with the binary cross-entropy objective (BCE baseline) and MFoM-F1 or MFoM-EER objectives. We train MFoM-base objectives either from "scratch" without weights pre-training or "tuning" the baseline weights. We compare results with our previous work [32].*

| Detectors | Performance measure is AvgEER (%) | | | | | |
|---|---|---|---|---|---|---|
| | BCE | MFoM-F1 | | MFoM-EER | | MFoM-F1 [32] |
| | | tuning | scratch | tuning | scratch | |
| manner | 11.65 | 10.67 | 11.17 | **10.61** | 10.73 | 13.40 |
| place | 16.94 | 14.30 | 17.59 | **14.23** | 14.56 | 15.67 |
| fmanner | 11.46 | 10.53 | 10.53 | 10.42 | **10.39** | 10.86 |
| fplace | 15.91 | 14.37 | 15.05 | **14.10** | 14.48 | 14.84 |

TABLE II

*Performance of place and manner attribute detectors per each class, comparison of the baseline models trained with binary cross-entropy and models trained with MFoM-EER (tuning). Total length of every attribute class is measured in minutes in the OGI-TS dataset [23]. Performance measure is EER (%), i.e., the lower the better.*

| Detectors | Attribute Classes | Total (min.) | Baseline | MFoM-EER tuning |
|---|---|---|---|---|
| Manner | Fricative | 54.51 | 12.20 | **11.12** |
| | Glides | 20.27 | 21.75 | **17.59** |
| | Nasal | 36.48 | 9.35 | **8.23** |
| | Stop | 56.81 | 12.66 | **12.15** |
| | Voiced | 246.64 | 8.88 | **8.62** |
| | Vowel | 165.05 | 9.21 | **9.01** |
| Place | Coronal | 117.33 | 24.35 | **22.41** |
| | Dental | 31.35 | 20.88 | **18.20** |
| | Glottal | 4.04 | 15.85 | **10.76** |
| | High | 52.12 | 16.08 | **15.25** |
| | Labial | 35.59 | 16.63 | **14.52** |
| | Low | 51.21 | 12.92 | **12.53** |
| | Middle | 70.35 | 17.57 | **17.24** |
| | Palatal | 10.90 | 12.98 | **11.11** |
| | Velar | 21.50 | 15.34 | **12.54** |

Interestingly, the MFoM-EER objective function with class-wise (macro) averaging seems to improve significantly the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2964953, IEEE/ACM Transactions on Audio, Speech, and Language Processing



recognition of rare classes, as shown in Table II. In fact, the recognition of the */glottal/* class, which has the smallest amount of training samples (4.04 minutes in the OGI-TS corpus), gains 5% absolute improvement in performance as compared with result obtained using a baseline neural architecture trained with binary cross-entropy. Conversely, it seems that the manner class despite having more training samples, namely */voiced/*, gained only a slight improvement, specifically from 8.88% to 8.62%.

We conclude this section highlighting some important configuration details:

- MFoM-based objectives (F1 and EER) are optimized with Adam [71], which is an adaptive learning rate algorithm, and a starting learning rate of 0.001.
- The averaging strategy is crucial. Class-wise MFoM averaging strategy over mini-batch allows to boost baseline performance, whereas, micro averaging does not improve significantly the baseline performance in any of the conditions (scratch or tuning).
- Experimenting with tanh, sigmoid, ReLU, and ELU as the output activation functions of the CRNN model showed us that tanh leads to the best performance.

The above-discussed configurations have been achieved using Bayesian optimization techniques [72], which allowed us to deploy MFoM-based training strategies from *scratch*, without pre-training the network parameters.

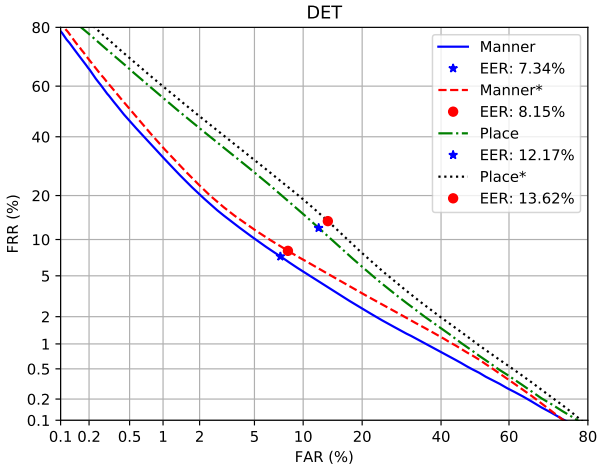


Fig. 7. Performance comparison of the proposed attribute detectors trained with MFoM-EER (*tuning*) versus detectors trained with MFoM-F1 (*tuning*) objective from the previous work [32], i.e., Manner/Place versus Manner*/Place* respectively.

### B. Attribute-based Features for Spoken Language Recognition

Using universal speech articulatory attributes, we assume that every language has its different quantitative content of speech attributes, i.e. distribution of attributes among languages. In Fig. 6 (on the left, manner attributes; on the right, place attributes), every color patch represents the mean value of the detection scores for the attribute classes. The mean values are calculated on the NIST LRE 2017 [65] dataset using attribute detection models trained on the OGI-TS dataset. It can be noticed that the most frequent manner attributes, detected in the NIST dataset, are *voiced* and *vowels*. The most diverse manner class across all languages is fricative. British English (*eng-gbr*) has the most amount of fricative sounds comparing other languages. Coronal and middle (*mid*) place attributes are classes with the most amount of detected observations in the NIST corpus. The amount of coronal sounds has the most variety from language to language. As such, we believe that both manner and placed properties might benefit spoken language recognition tasks. Since the goal of the present work is to demonstrate the complementarity of speech attributes to the acoustic features, we stack those attributes with the *basic speech features* (e.g., 80 BNF + 9 *fuse-place* = 89-dim), namely MSpec, SDC, or BNFs, and form six different feature combination solutions, as shown in Table III. Next, we apply *singular value decomposition* (SVD) and reduce dimension to 80-dimensional feature vectors in order to keep the system complexity comparable across different configurations. We have thus obtained *attribute-based features*, which are employed to generate i-vectors as discussed in Section VI-B3.

### C. Spoken Language Recognition Analysis

TABLE III
*The results of the spoken language recognition (SLR) system using bottleneck features (BNF), mel-spectrogram (MSpec), shifted delta cepstral (SDC) features and speech attribute features (manner, place, fusion manner and fusion place, see Fig. 5). Performance measures are F1 and Cavg.*

| Features | F1, % | | Cavg., % | |
|---|---|---|---|---|
| | dev-set | eval-set | dev-set | eval-set |
| BNF | 79.2 | 77.6 | 23.8 | 25.1 |
| [BNF, place] | 79.8 | 77.9 | 22.6 | 24.6 |
| [BNF, manner] | **80.8** | 77.9 | **22.3** | 24.6 |
| [BNF, place, manner] | 80.3 | 78.1 | 22.5 | 24.6 |
| [BNF, fplace] | 79.9 | **78.1** | 22.9 | **24.1** |
| [BNF, fmanner] | 80.0 | 77.7 | 22.7 | 24.9 |
| [BNF, fplace, fmanner] | 80.0 | 78.0 | 22.4 | 24.6 |
| MSpec | 58.8 | 56.6 | 47.1 | 48.4 |
| [MSpec, place] | **63.3** | 59.7 | **41.2** | 44.6 |
| [MSpec, manner] | 60.5 | 58.2 | 45.1 | 46.2 |
| [MSpec, place, manner] | 63.1 | 60.1 | 42.3 | **43.6** |
| [MSpec, fplace] | 62.4 | **60.5** | 42.4 | 44.0 |
| [MSpec, fmanner] | 61.5 | 59.5 | 43.7 | 45.2 |
| [MSpec, fplace, fmanner] | 61.9 | 60.2 | 42.9 | 44.5 |
| SDC | 61.1 | 58.8 | 44.2 | 46.0 |
| [SDC, place] | 63.2 | 61.7 | 42.5 | 42.5 |
| [SDC, manner] | 59.9 | 60.2 | 45.0 | 44.2 |
| [SDC, place, manner] | 62.8 | 61.6 | 42.7 | 42.8 |
| [SDC, fplace] | **65.5** | **63.2** | **39.9** | **41.4** |
| [SDC, fmanner] | 61.9 | 61.3 | 43.1 | 43.2 |
| [SDC, fplace, fmanner] | 60.9 | 61.0 | 44.5 | 44.2 |

In this section, we confirm the positive effect of the phonetic BNF features on the baseline i-vector systems. Later, we compare the contribution of the proposed multi-lingual attribute features incorporated in the baseline systems.

*1) Baseline:* We conduct SLR experiments on the NIST LRE 2017 task. As previously mentioned (in Section VI-B), we built three different baseline SLR systems based on three

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2964953, IEEE/ACM Transactions on Audio, Speech, and Language Processing



different features: MSpec, SDC, and the deep bottleneck features (BNF). The BNF baseline system was trained on English data only (*Switchboard-1* and *Fisher* corpora, approx. 2000 hours). Phonetic domain information trained with BNF significantly contributes to SLR systems, comparing to non-phonetic MSpec and SDC systems. In Table III we see that BNF achieves lower $C_{avg}$ than both MSpec and SDC. Overall, BNF strikingly outperforms MSpec by about 73% and SDC by about 65% relative on evaluation dataset.

*2) Effect of Attribute Features on SLR:* The proposed technique expands the SLR baseline configurations by injecting speech attribute information extracted with a bank of detectors implemented as discussed in Section VII-A. We obtained six additional LRE systems for each of the three baseline SLR systems, namely: *manner*, *place*, *fuse-manner* (fmanner), *fuse-place* (fplace) and combinations. Independently of whether *mel-spectrogram* (MSpec) or *shifted delta cepstral* (SDC) features are selected, we have witnessed a consistent performance gain in the SLR when leveraging articulatory attributes, i.e., a beneficial overall effect on the automatic language discrimination is achieved by combining standard features and attributes. The performance of the BNF-based system was also slightly improved by exploiting additional information at attribute level: the F1 score was raised from 77.6% up to 78.1% along with a 3% relative improvement in terms of Cavg. Moreover, place of articulation features appear to be more diverse across languages (see Fig. 6 (right)), since the mean values of place scores are significantly varying from language to language, which is not observable for the manner of articulation scores. As a consequence, place attributes improve overall language recognition and boost the performance of both systems: for the SDC system the F1 measure is increased from 58.8% up to 63.2%, while for the MSpec system F1 score increases from 56.6% to 60.5%. Moreover, from Tables I and III, we noticed improvements on place of articulation detector cascade as well as improvements in spoken language recognition. It seems that spoken language recognition performance is boosted when moving from the stand-alone *place* to the fusion-place (*fplace*) configuration. The SDC-based micro-F1 goes from 61.7% to 63.2%, the MSpec-based micro-F1 increases from 59.7% to 60.5%, and the BNF micro-F1 goes from 77.9% up to 78.1%. On the other hand, moving from manner attributes to fusion-manner, it improves systems based only on spectral (MSpec) and SDC features.

## VIII. Conclusions

This paper contributes to the front-end study of the spoken language recognition (LRE) pipeline. It combines the knowledge gained from our previous work with the maximal figure-of-merit mathematical framework (MFoM), multi-label acoustic event detection, and speech articulatory features into a single framework. We show that manner and place of articulation features (speech attributes) jointly modeled and extracted at the output of a deep model provide a parsimonious representation of any spoken language; furthermore, we can train attribute detectors on a relatively small dataset (7 hours) compared with the large amount of training material, namely Switchboard dataset (2000 hours) for the BNF features. In addition, attribute feature scores correspond to universal phonetic cues that can be used to describe any spoken language.

Finally, we show that the proposed maximal figure-of-merit (MFoM) learning approach directly embeds micro-F1 and EER performance measures into backpropagation optimization. This allows us to encode multi-label information of multiple speech attribute classes into a "units-vs-zeros" misclassification measure to be used directly in the MFoM framework. MFoM allows us to approximate the metric of interest with a differentiable function, so that gradient-based optimization algorithms can be applied to learn the DNN parameters. Experimental evidence demonstrates that the proposed optimization strategy outperforms that based on more conventional binary cross-entropy objective function. Furthermore, by applying Bayesian optimization techniques we managed to find hyperparameters of neural network appropriate to train MFoM objectives from scratch, without any initial weights pre-training.

## IX. Acknowledgements

This research was partially funded by the Academy of Finland (grant #313970), Finnish Scientific Advisory Board for Defence (MATINE) project #2500M-0106 and the ARAP grant from the Institute for Infocomm Research, A*STAR. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

## References

[1] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 623–666, 1948.

[2] H. Li, B. Ma, and K.-A. Lee, "Spoken language recognition: From fundamentals to practice." *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.

[3] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, pp. 210–229, 2006.

[4] M. Zissman and T. Gleason, "Automatic dialect identification of extemporaneous conversational latin american spanish speech," in *ICASSP*, 1995, pp. 777–780.

[5] P. Matějka, P. Schwarz, J. Cernocký, and P. Chytil, "Phonotactic language identification using high quality phoneme recognition," in *INTERSPEECH*, 2005.

[6] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, pp. 788–798, 2011.

[7] K. A. Lee, H. Li, L. Deng, V. Hautamki, W. Rao, X. Xiao, A. Larcher, H. Sun, T. Nguyen, G. Wang, A. Sizov, J. Chen, I. Kukanov, A. Poorjam, T. Trong, C.-L. Xu, H.-H. Xu, B. Ma, E.-S. Chng, and S. Meignier, "The 2015 NIST language recognition evaluation: the shared view of I2R, Fantastic4 and SingaMS," in *INTERSPEECH*, 2016, pp. 3211–3215.

[8] T. N. Trong, V. Hautamäki, and K. A. Lee, "Deep language: a comprehensive deep learning approach to end-to-end language recognition," in *Proc. Odyssey*, 2016.

[9] B. Jiang, Y. Song, S. Wei, J.-H. Liu, I. V. McLoughlin, and L.-R. Dai, "Deep bottleneck features for spoken language identification," *PLoS ONE*, vol. 9, 2014.

[10] P. Matějka, L. Zhang, T. Ng, S. H. Mallidi, O. Glembek, J. Ma, and B. Zhang, "Neural network bottleneck features for language identification," in *Proc. Odyssey*, 2014, pp. 299–304.

[11] X. Huang and M. Jack, "Semi-continuous hidden markov models for speech signals," *Computer Speech & Language*, vol. 3, no. 3, pp. 239–251, jul 1989.

[12] M. McLaren, L. Ferrer, and A. Lawson, "Exploring the role of phonetic bottleneck features for speaker and language recognition," *ICASSP*, pp. 5575–5579, 2016.

[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] 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.

[15] 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.

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

[17] W. Hu, Y. Qian, F. K. Soong, and Y. Wanga, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154–166, 2015.

[18] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 333–353, October 2000.

[19] S. M. Siniscalchi and C.-H. Lee, "A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition," *Speech Communication*, vol. 51, pp. 1139–1153, 2009.

[20] H. Behravan, H. V, S. M. Siniscalchi, T. Kinnunen, and C. Lee, "i-vector modeling of speech attributes for automatic foreign accent recognition," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 24, no. 1, pp. 29–41, 2016.

[21] T. Poggio and F. Girosi, "Regularization algorithms for learning that are equivalent to multilayer networks," *Science*, vol. 247, pp. 978–982, 1990.

[22] M.-L. Zhang, "Multilabel neural networks with applications to functional genomics and text categorization," *KDE*, vol. 18, no. 10, pp. 1338–1351, Oct. 2006. [Online]. Available: http://dx.doi.org/10.1109/tkde.2006.162

[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. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.40.7091

[24] J. Nam, J. Kim, E. Loza Mencía, I. Gurevych, and J. Fürnkranz, "Large-scale multi-label text classification - revisiting neural networks," in *ECML-PKDD-14*. Springer Berlin Heidelberg, 2014, pp. 437–452.

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

[26] S. Katagiri, C.-H. Lee, and B.-H. Juang, "New discriminative training algorithms based on the generalized probabilistic descent method," in *Neural Networks for Signal Processing Proceedings of the 1991 IEEE Workshop*.

[27] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

[28] S. Katagiri, B.-H. Juang, and C.-H. Lee, "Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2345–2373, 1998.

[29] M. G. Rahim and C. Lee, "String-based minimum verification error (SB-MVE) training for speech recognition," *Computer Speech & Language*, vol. 11, no. 2, pp. 147–160, 1997. [Online]. Available: https://doi.org/10.1006/csla.1997.0026

[30] 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 Trans. on Inf. Syst.*, vol. 24, 2006.

[31] 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 *ICASSP 2014*.

[32] I. Kukanov, V. Hautamäki, S. M. Siniscalchi, and K. Li, "Deep learning with maximal figure-of-merit cost to advance multi-label speech attribute detection," in *SLT 2016, San Diego, CA, USA*, pp. 489–495.

[33] I. Kukanov, V. Hautamki, and K. A. Lee, "Recurrent neural network and maximal figure of merit for acoustic event detection," DCASE2017 Challenge, Tech. Rep., September 2017.

[34] I. Kukanov, V. Hautamäki, and K. A. Lee, "Maximal figure-of-merit embedding for multi-label audio classification," in *ICASSP 2018, Calgary, Alberta, Canada*.

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

[36] 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, August 27-31, 2007*, 2007, pp. 1829–1832.

[37] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein, D. Byrd, E. Bresch, P. Ghosh, A. Katsamanis, and M. Proctor, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1307–1311, sep 2014.

[38] C. Lopes and F. Perdigao, "Phoneme recognition on the TIMIT database," in *Speech Technologies*. InTech, jun 2011. [Online]. Available: http://dx.doi.org/10.5772/17600

[39] 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.

[40] M. Ashby and J. Maidment, *Introducing Phonetic Science*, ser. Cambridge Introductions to Language and Linguistics. Cambridge University Press, 2005.

[41] S. M. Siniscalchi, T. Svendsen, and C.-H. Lee, "Toward a detector-based universal phone recognition," *IEEE*, 2008.

[42] F. Katamba, *An Introduction to Phonology*. Longman Pub Group, 1989.

[43] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions On Audio, Speech, and Language Processing*, October 2014. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=244766

[44] T. Virtanen, A. Mesaros, T. Heittola, M. Plumbley, P. Foster, E. Benetos, and M. Lagrange, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*. Tampere University of Technology. Department of Signal Processing, 2016.

[45] B. H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Transactions on Signal Processing*, vol. 40, no. 12, pp. 3043–3054, Dec. 1992.

[46] K. Plunkett and J. Elman, *Exercises in Rethinking Innateness: A Handbook for Connectionist Simulations*. MIT Press, 1997.

[47] J. Xu, "A weighted linear discriminant analysis framework for multi-label feature extraction," *Neurocomputing*, vol. 275, pp. 107–120, Jan. 2018.

[48] M.-J. Zhao, N. U. Edakunni, A. C. Pocock, and G. Brown, "Beyond Fano's inequality: bounds on the optimal F-score, BER, and cost-sensitive risk and their implications." *J. Mach. Learn. Res.*, vol. 14, 2013.

[49] R. Linsker, "Towards an organizing principle for a layered perceptual network," in *Neural Information Processing Systems*. American Institute of Physics, 1988, pp. 485–494.

[50] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, Aug. 2018.

[51] B. Hu and W. Dong, "A study on cost behaviors of binary classification measures in class-imbalanced problems," *CoRR*, vol. abs/1403.7100, 2014.

[52] E. Eban, M. Schain, A. Mackey, A. Gordon, R. Rifkin, and G. Elidan, "Scalable Learning of Non-Decomposable Objectives," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Singh and J. Zhu, Eds., vol. 54. Fort Lauderdale, FL, USA: PMLR, 2017, pp. 832–840.

[53] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[54] V. Mingote, A. Miguel, D. Ribas, A. Ortega, and E. Lleida, "Optimization of false acceptance/rejection rates and decision threshold for end-to-end text-dependent speaker verification systems," *INTERSPEECH*, 2019.

[55] A. Sanyal, P. Kumar, P. Kar, S. Chawla, and F. Sebastiani, "Optimizing non-decomposable measures with deep networks," *Mach. Learn.*, vol. 107, no. 8-10, pp. 1597–1620, 2018.

[56] V. M. Tikhomirov, "On the notion of mean," in *Selected Works of A. N. Kolmogorov*. Springer Netherlands, 1991, pp. 144–146.

[57] Y. H. Hu, *Handbook of Neural Network Signal Processing*. Boca Raton, FL, USA: CRC Press, Inc., 2000.

[58] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, 2015.

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

[60] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[61] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, Y. W. Teh and M. Titterington, Eds., vol. 9. PMLR, 13–15 May 2010, pp. 249–256.

[62] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, 2016.

[63] E. Çakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM TASLP*, vol. 25, no. 6, pp. 1291–1303, jun 2017.

[64] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)." *CoRR*, 2015.

[65] NIST, "Nist 2017 language recognition evaluation plan," 2017. [Online]. Available: https://www.nist.gov/itl/iad/mig/nist-2017-language-recognition-evaluation

[66] K. A. Lee, H. Li, L. Deng, V. Hautamaki, and et .al, "The 2015 nist language recognition evaluation: the shared view of i2r, fantastic4 and singams," *Interspeech*, 2016.

[67] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, pp. 210–229, 2006.

[68] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for svm-based speaker recognition." in *INTERSPEECH*. ISCA, 2006.

[69] D. A. Leeuwen and N. Brümmer, "Speaker classification I," C. Müller, Ed. Berlin, Heidelberg: Springer-Verlag, 2007, ch. An Introduction to Application-Independent Evaluation of Speaker Recognition Systems, pp. 330–353.

[70] N. Brümmer, "Measuring, refining and calibrating speaker and language information extracted from speech," 2010.

[71] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization." *CoRR*, vol. abs/1412.6980, 2014.

[72] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2011, pp. 2546–2554. [Online]. Available: http://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf

[73] D. Rumelhart, G. Hinton, and R. J. Williams, "Learning internal representations by error propagation," *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 1, 1986.

[74] E. Bottou, "Stochastic gradient learning in neural networks."

[75] M. Ranjbar, T. Lan, Y. Wang, S. N. Robinovitch, Z.-N. Li, and G. Mori, "Optimizing nondecomposable loss functions in structured prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 4, pp. 911–924, apr 2013. [Online]. Available: https://doi.org/10.1109/tpami.2012.168

[76] T. Joachims, "A support vector method for multivariate performance measures," in *Proceedings of the 22Nd International Conference on Machine Learning*, ser. ICML '05, 2005, pp. 377–384.


**Ivan Kukanov** received his B.Sc. degree with honors and M.Sc. degree in Applied Mathematics and Computer Science from the Saint-Petersburg State University, Russia in 2012 and 2014, respectively. He received his M.Sc. degree in Computer Science from the University of Eastern Finland, in 2015. From 2017 to 2019, he had been working as the research assistant of the ARAP program at I2R, A*STAR, Singapore. Since 2019, he is a scientist at I2R, A*STAR, Singapore.

**Trung Ngo Trong** is a Ph.D. candidate at the University of Eastern Finland (UEF), working on semi-supervised learning algorithm applied to signal processing. He graduated from Hanoi University of Science and Technology, with bachelor degree in computer engineering, and received his master in information technology from UEF. He received best paper award at IWSDS 2018.

**Ville Hautamäki** is a senior researcher, received the M.Sc. degree in Computer Science from the University of Eastern Finland, Finland in 2005. He received the Ph.D. degree in Computer Science from the same university in 2008. He worked from 2009 to 2011 as a research fellow at the Institute for Infocomm Research, A*STAR, Singapore. In 2013 he was a visiting scholar at the Georgia Institute of Technology, USA. He serves as an associate editor in Digital Speech Processing and is member of the IEEE SLT Committee.

**Sabato Marco Siniscalchi** is a Professor at the University of Enna, and affiliated with the Georgia Institute of Technology (Ga Tech). He received Doctorate degrees in Computer Engineering from the University of Palermo in 2006. In 2006, he was a Post Doctoral Fellow at the Ga Tech. From 2007 to 2010, he joined NTNU, Norway, as a Research Scientist. From 2010 to 2015, he was an Assistant Professor, first, and an Associate Professor, after, at the Kore University. From 2017 to 2018, he was a Senior Speech Researcher at Siri Speech Group, Apple Inc., Cupertino CA, USA. He acted as an associate editor in the IEEE/ACM Transactions on Audio, Speech and Language Processing, from 2015 to 2019. Dr. Siniscalchi is an elected member of the IEEE SLT Committee (2019-2021).

**Valerio M. Salerno** is an Assistant Professor at the Kore University of Enna, Italy. He received his Bachelor degree in Telematic Engineering from University of Catania, Italy, and both his Masters degree (cum Laude) in Telematic Engineering and Ph.D. from Kore University of Enna, Italy.

**Kong Aik Lee** (M'05-SM'16) is currently a Senior Principal Researcher at the Biometrics Research Laboratories, NEC Corp., Japan. He received his Ph.D. degree from Nanyang Technological University, Singapore, in 2006. From 2006 to 2018, he was a Scientist at the Human Language Technology department, $I^2R$, A*STAR, Singapore, where he led the speaker recognition group. He was the recipient of Singapore IES Prestigious Engineering Achievement Award 2013 for his contribution to voice biometrics technology. He serves as an Editorial Board Member for Elsevier Computer Speech and Language, and an Associate Editor for IEEE/ACM Transactions on Audio, Speech and Language Processing. He is an elected member of IEEE STL Committee.