AUTOMATIC DETECTION OF SPEECH UNDER COLD USING DISCRIMINATIVE AUTOENCODERS AND STRENGTH MODELING WITH MULTIPLE SUB-DICTIONARY GENERATION

Yi-Ying Kao[1], Hsiang-Ping Hsu[1], Chien-Feng Liao[2], Yu Tsao[2], Hao-Chun Yang[3], Jeng-Lin Li[3], Chi-Chun Lee[3], Hung-Shin Lee[4] and Hsin-Min Wang[4]

[1]Ministry of Justice, Investigation Bureau, R.O.C.
[2]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[3]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[4]Institute of Information Science, Academia Sinica, Taiwan

ABSTRACT

In this paper we aim to tackle the *Cold* sub-challenge proposed in the INTERSPEECH 2017 ComParE Challenge. The goal is to determine whether given speech is under cold condition. In this paper we present two frameworks. One of them is based on an alternative neural network-based autoencoder using two different loss functions. The first one is the standard reconstruction error used in unsupervised autoencoder, and the hinge loss (second loss function) is incorporated into the middle layer to attract utterances spoken by the same condition into similar identity code spaces. The classification is then carried out by comparing the cosine similarity of identity codes between the target and the mean of cold and non-cold utterances. With a simple logistic regression combining our method and the baseline systems predictions, we achieve 65.81% and 66% UAR on development set and test set provided by 2017 ComParE, respectively. Another approach is based on strength modeling, where diverse classifiers' confidence outputs are concatenated to original feature space as input to the support vector machine. The feature representations are derived from multiple sub-dictionary within the framework of GMM Fisher-vector encoding and eGeMAPS functional features concatenating with diverse classifiers. We achieve 70.2% and 65.5% on development and test set provided by 2017 ComPareE, respectively.

Index Terms— cold detection, discriminative autoencoders, deep neural networks, computational paralinguistics

1. INTRODUCTION

The INTERSPEECH ComParE Challenge series has been introducing various of unseen problems in the computational paralinguistic field with well-defined protocol and competitive benchmark, such as emotion and gender classification [1]. The results can immensely benefit diverse application domains, assisting our daily life demands from medical diagnosis (detection of parkinson's condition [2]) to law enforcement applications (lie detection [3]) . In this paper we propose two frameworks aiming to solve the *Cold* sub-challenge proposed in the INTERSPEECH 2017 ComParE[4], given a corpus consists of speech under UPPER RESPIRATORY TRACT INFECTION CORPUS (URTIC) condition, normal health condition needs to be determined. The discriminative autoencoder framework is motivated by the work done in speaker verification task [5]. The autoencoder is a symmetric neural network that is trained to reconstruct its input at the output layer through an unsupervised learning fashion and can learn the most salient and useful representation of the data [6]. It has been widely applied to many speech processing tasks, including but not limited to speech enhancement [7], speech dereverberation [8, 9], and reverberant speech recognition [10]. In our work not only the output reconstruction is considered but also focus on the middlemost layer where identity codes (i-codes) are produced, desiring to attract utterances spoken by the same condition into similar identity code spaces, using a modified hinge-like error function.

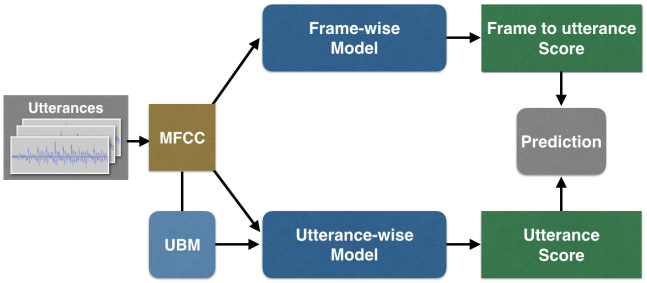

Fig. 1. Proposed framework with extracting features and training discriminative model.

Another strategy to enhance the recognition and robustness is to use a fusion scheme. The combination of complementary modalities offers important information for sophisticated classification tasks. The concept of sub-dictionary inspired by [11] is used for discriminative sub-feature sets generation that could be incorporated using the state-of-the-art fusion strategy. In addition to the conventional fusion approaches, a hierarchical feature fusion framework was proposed in [12] which exploits not only the feature level but decision level with discriminative power. Recently, a newly proposed scheme, strength model, provides another fusion perspective for further improvement [13]. A multiple-classifier structure shows that each different subset of features could be combined with respect to their corresponding optimal classifier [14]. Hence, we integrate the GMM sub-dictionary fisher-vector and eGeMAPS functional features by means of the strength model and optimize each sub-dictionary and feature set with different classifiers.

The remainder of this paper is organized as follows. In Section 2 and 3 we introduce the proposed discriminative autoencoders with detailed explanation of the objective functions and strength modeling

978-1-5386-8151-0/18/$31.00 ©2018 IEEE

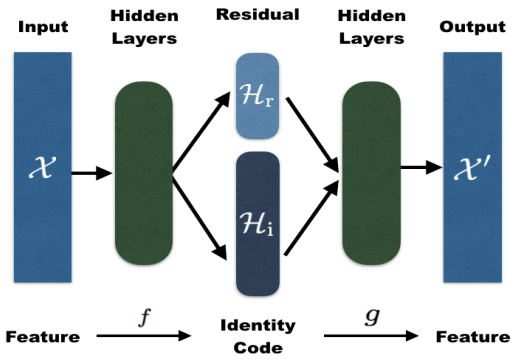

Fig. 2. Discriminative autoencoder architecture.

with various learning and fusion methods. The experimental results are given in Section 4, Section 5 concludes with discussions and future prospect.

2. DISCRIMINATIVE AUTOENCODER

To train the model we explore two ways which are illustrated in Figure 1. First we exploit the frame-wise approach by extracting MFCC features and train the model frame by frame with reconstruction loss and hinge loss function. The prediction is then made by simply compare the mean of i-codes within an utterance to the total average of cold utterances and non-cold utterances, respectively. Another approach uses i-vectors [15], a popular method in the field of speaker recognition which represents variable-length speech signals by fixed-length tokens while preserving the speaker characteristics. The rest of training and scoring procedure is the same as frame-wise approach. The reconstruction function Eq 1 and hinge loss function Eq 2 are applied to the $\mathcal{X}'$ and $\mathcal{H}_\mathrm{i}$ illustrated in Fig 2, respectively.

2.1. The reconstruction error

Our proposed discriminative model contains a pair of learned transformation functions $f(\cdot)$ and $g(\cdot)$ as encoder and decoder, respectively. The encoder projects input to a latent space $\mathcal{H}$, and the decoder tries to reconstruct the input from the latent space codes. Given a set of training data $\mathcal{X}$ the model will go through the reconstruction procedure $\mathcal{X} \xrightarrow{f} \mathcal{H} \xrightarrow{g} \mathcal{X}'$, the average reconstruction error comes from the residual sum of squares between inputs $x \in \mathcal{X}$ and its reconstructed outputs $x' = g(f(x))$, denoted in Eq 1 where $|\mathcal{X}|$ is the sample size.

$$\mathcal{L}_\mathrm{r}(x, x') = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \|x - x'\|^2 \tag{1}$$

2.2. The hinge loss function

We then focus on the middlemost layer where $\mathcal{H}_\mathrm{i}$ and $\mathcal{H}_\mathrm{r}$ lie within. The $\mathcal{H}_\mathrm{r}$ part contains content other than the target information we desired, hence this residual part is not in our concern. The objective of our loss function is to attract utterances that belong to the same class into similar region in the code space, where similarity is measured by the inner product of two $\mathcal{H}_\mathrm{i}$ codes. Moreoever, we only target at the pairs that have ambiguous similarity, and the hinge-style margin threshold is therefore introduced to the loss function. For positive pairs only the ones that have inner products below the positive margin will contribute to the error otherwise they are set to zero, and vice versa.

$$\mathcal{L}_\mathrm{h}(\mathcal{H}_\mathrm{i}) = \frac{1}{|\mathcal{H}_\mathrm{i}|} \sum_{h, h' \in \mathcal{H}_\mathrm{i}} (max(0, \mathrm{m}_p - h) + max(0, h' - \mathrm{m}_n))^2 \tag{2}$$

where $|\mathcal{H}_\mathrm{i}|$ is the sample size, h and h' are the inner products of i-code pairs that belong to the same class and different class respectively, m_p and m_n is the margin size for identical pairs and distinct pairs.

3. STRENGTH MODELING

3.1. Sub-Dictionary Learning

A sub-dictionary strategy aims at expanding feature space with categorical discriminative power. Conventionally, the dictionary learning method mainly depends on the unsupervised learning using low level descriptors trained on the entire training set. Detailed discriminative characteristics would be neglected when considering such a general scheme of representational learning with respect to all the samples. Thus, we specify sub-dictionaries with pre-defined criteria , including label-derived and unsupervised-clustered approaches. For our binary classification task, two categories containing Cold and No-Cold are specified referring to the true label. Additionally, k-means as an unsupervised clustering to provide another mean for splitting the data into two categories for sub-dictionary learning. We adopt GMM as probability distribution based dictionary with its capability of soft clustering assignment. In the end, we derive a general GMM, cold-specific GMM, no-cold-specific GMM, and two unsupervised-specific GMMs.

3.2. Fisher Encoding

Fisher vector, originated from image recognition community, also shows competitive accuracies in many computational paralinguistics analyses tasks [16]. The use of GMM-FV encoding has the advantage of being both a generative and discriminative model. If the low-level acoustic feature set is denoted by $X = \{x_t, t = 1...T\}$ with D dimensions and the set of parameters of GMM is $\lambda = \{w_i, \mu_i, \Sigma_i, i = 1...K\}$ where w_i , μ_i and Σ_i are correspond to the zeroth (weight), first (mean vector) and second (covariance matrix) statistics for each mixture of Gaussian, respectively. Then, the fisher vector is formed by concatenating the gradient of first and second derivatives and neglect the zeroth moment. Hence, the final feature is a $2 \times K \times D$ super vector.

3.3. Fusion

3.3.1. Strength Model

A novel hierarchical fusion framework, strength model, is proposed to jointly obtain the potential of feature level and decision level fusion. The effectiveness and robustness of using strength model has been shown to achieve the state-of-the-art performances. [13]. Methodologically, the decision scores from multiple modalities will be stacked up with multiple feature descriptors which act as a new input of the other classifier[17]. To compute decision scores, we use four well-known classifiers, namely support vector machine, adaboost, random forest, and naive bayes, anticipating that the various classifiers can provide different perspectives to the classification

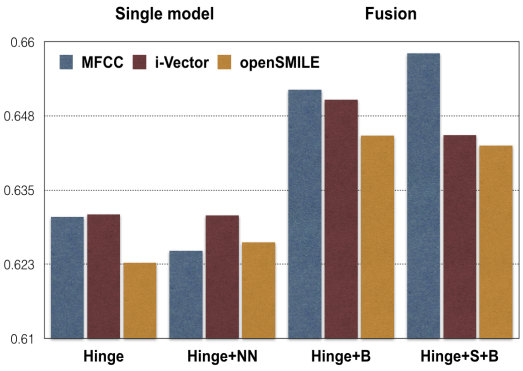

Fig. 3. Experimental results containing single models and fusion combinations.

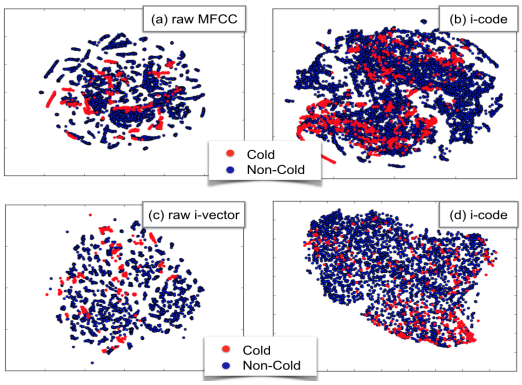

Fig. 4. Scatter plot of the results by t-SNE for MFCC and i-vector features.

problem. Then multiple feature vectors trained by using different specific GMMs are jointly concatenated constituting our final feature.

4. EXPERIMENTS AND RESULTS

Since the official dataset contains 9505 training data and 9596 developing data with highly unbalanced ratio between Cold(C) and No-cold(NC) instances, for the discriminative model training, we repeatedly copy the C data to match the amount of NC data since the neural network favors a larger training data. As for the strength modeling, we downsample the NC data randomly to amount of cold data for feasible and effective experiment conduction. For ease of reproductivity, all the results are implemented by open-source toolkits. The low-level descriptors (LLD) are extracted through standard openSMILE [18] configurations. Meanwhile, we use scikit-learn [19] for GMM modeling and other classifiers might be applied in the following works. The metric used for assessing the Cold sub-challenge is Unweighted Average Recall (i.e. mean recall on both classes).

4.1. Discriminative Autoencoder

4.1.1. Experimental setup

For the frame-wise approach, speech parameters were represented by a 60 dimensional feature vector of Mel-frequency cepstral coefficients (MFCCs) extracted with open-source toolkit librosa [20], with the frame length of 128 ms and the frame shift of 32 ms. For the utterance-wise approach, an UBM consisting of 2,048 Gaussian components with diagonal covariance matrices, the total variability model (i-vectors) with rank 600, were trained on the official training dataset by open-source toolkit Kaldi [21]. Additionally, we also incorporate 6373 dimensional openSMILE feature vector used by the official baseline for comparison.

For model that achieves the best score on developing set, the number of nodes in the i-code and noise layer are both 100. There are two hidden layers each with 256 nodes. We initialize the weights with the Glorot uniformly distribution which is fit for the tanh activation function [22]. For the optimizer we use adaptive estimates of lower-order moments algorithm Adam to update the model parameters [23], with the initial learning rate of 1e-4.

4.1.2. Results

We compare the result of three features on the discriminative model, denoted as "Hinge" in Figure 3, along with further experiments of two additional methods and a fusion approach. For the first method, denoted as "Hinge+NN", a fully-connected neural network with softmax output layer is trained with the i-codes extracted by the model and directly predicts the utterance labels. We can see some improvement on openSMILE feature. Secondly, we swap the i-code layer with a two-nodes softmax layer, which directly output class predictions. For this approach we need further study on the results. Finally we concatenate the probability outputs of various methods including baseline system given by the organizers to feature vectors and train a simple logistic regressor to make the final prediction. Two combinations are denoted as "Hinge+B" as Hinge+NN fusion with two baseline system and Hinge+S+B", as we combine the second method mentioned above with "Hinge+B" fusion.

Finally, to demonstrate the effectiveness of the proposed model, we visualize the i-codes of MFCC and i-vector on developing set be embedded into a 2-dimensional plane by t-Distributed Stochastic Neighbor Embedding (t-SNE) in Figure 4, where (a)(c) and (b)(d) show the raw input features and i-codes, respectively.

4.2. Sub-Dictionary Strength Modeling

4.2.1. Feature Set Comparison

In this experiment, we evaluate the performance of ComParE16, eGeMAPs in this task. Both feature sets have been preprocessed by feature-level z-score normalization beforehand, and then the standard support vector machine is used as the final binary classifier. From table 2, we observed that although with much fewer dimensions than Compare16, eGeMAPs can still achieve competitive accuracy on the dataset. However, Compare2016 still achieves the highest recognition rate, and hence, the following discussions will be based on this feature set.

4.2.2. Sub-Dictionary baseline results

To evaluate discriminative capability in this binary classification task for each sub-dictionaries, we take each at a time for Fisher-vector encodings. From table 3, we can conclude with some observations: first, all sub-dictionaries outperform the general dataset. Since URTIC is a large and complex database composed by many different people, the method we proposed provides a more detailed view

Table 1. Fusion results: Results of different fusion schemes on 2000 balance sampled development set; Gen: General, C:Cold, NC: No Cold, eGs: eGeMAPs, usvA: Unsupervised cluster 1, usvB:Unsupervised cluster 2, Ada: Adaboost; Feature levels and decision levels are separated by a single slash in strength modeling.

Top1	Top2	Top3	Top4	Top5
Decision Level Fusion				
0.667 usvB(64,Ada)+ usvA(32,Ada)	0.664 usvB(64,Ada)+ usvB(64,SVM)+ usvA(32,Ada)	0.664 usvB(64,Ada)+ usvA(32,Ada)	0.663 usvB(64,Ada)	0.661 usvB(64,SVM)
Feature Level Fusion				
0.684 gen(32)+C(16)+ usvB(32)+eGs	0.683 NC(32)+usvB(16)+ eGs	0.682 C(16)+usvB(32)+ eGs	0.681 usvA(16)+usvB(32)+ eGs	0.677 gen(64)+C(16)+ usv(32)+eGs
Strength Model				
0.702 gen(32)+C(16)+usvB(32)/ gen(64,Ada)+C(16,Ada)+ usvA(32,Ada)+ usvB(64,Ada)+ eGs	0.700 gen(32)+C(16)+usvB(32)/ usvB(16,Ada)+ usvB(64,Ada)+eGs	0.699 gen(32)+C(16)+usvB(32)/ usvA(32,Ada)+ usvB(64,Ada)+eGs	0.697 gen(32)+C(16)+usvB(32)/ gen(16,Ada)+ usvA(32,Ada)+ usvB(32,Ada)+eGs	0.697 gen(32)+C(16)+usvB(32)/ gen(16,Ada)+ usvA(32,Ada)+ usvB(64,Ada)+eGs

Table 2. Baselines system: Comparison of feature sets under different SVM parameters; C: Complexity parameter of SVM

C	ComPare_16	eGeMAPs
10	0.591	0.598
1	0.591	0.604
10_{-1}	0.591	0.612
10_{-2}	0.590	0.604
10_{-3}	0.600	0.624
10_{-4}	0.623	0.623
10_{-5}	0.629	0.613
10_{-6}	0.609	0.613

on the specific characteristics of the dataset. Second, Cold sub-dictionary is better than No-Cold. It may be due to the fact that the patterns of voice characteristics for cold people are more distinctive, and our sub-dictionary takes advantage of this in improving the results. In the last, we observe that using unsupervised clustering method actually achieves the highest recognition accuracy. This implies that the perspective provided as output of an unsupervised clustering may possess a different information on the data compared to using just the labels for sub-dictionary training in this task.

4.2.3. Fusion Scheme

On the purpose of integrating information of distinct modalities, we evaluate both decision level fusion and feature level fusion in the experiments. A hierarchical and logistic design is used to obtain the optimal fusion strategy. As illustrated in Table 1, the top five outcomes decision scores are shown. The unsupervised cluster 2 (usvB) dominates the decision level results. Feature level fusion provides another insight into the integration of these five dictionaries and functional descriptors. Mixture numbers of 32 and 64 are used for general dictionary learning and 16, 32 for the rest. Different combinations can all achieve 0.68 UAR on the development set. It shows that feature level fusion is more advantageous than decision level fusion.

Table 3. Sub-dictionary baselines: The best prediction results for each sub dictionary only encoding; General:whole data set without segmentation, usvA:unsupervised cluster A, unsupervised cluster B

General	Cold	No-Cold	usvA	usvB
0.659	0.666	0.663	0.67	0.686

Finally, by means of the decision score fusion scheme, the most favorable classifier will be selected from each dictionary to obtain the decision score to be used in the strength modeling. Thus, we specify the feature combination to be used in the strength model, i.e., gen(32)+C(16)+usvB(32)+eGs, as a consequence from our empirical observation that achieves the highest performance in feature level fusion. On the basis of the pre-specified combination, we examine the effectiveness of decision scores with appropriate classifier and corresponding optimal parameters. The best accuracy achieved on the development set is 70.2%, and the test set accuracy is 65.5%.

5. CONCLUSIONS

In this task, given the data with no more information other than the class labels, the problem can be challenging. With our proposed discriminative autoencoder, a simple light-weight neural network with 60 dimensional MFCC feature can achieve baseline level accuracy on the developing set, with more time on fine-tuning the model parameters. The framework demonstrates promising potential. Besides, the strength model along with the sub-dictionary technique and eGeMAPS feature set introduces a distinctive insight and convincing results. In the future, an optimistic performance can be anticipated by examining feasible features and sophisticated sub-dictionary partition criterion.

6. ACKNOWLEDGEMENTS

The authors would like to thank the financial support provided by Ministry of Justice (Grant number:107-1301-05-17-02)

7. REFERENCES

[1] Björn W Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian A Müller, Shrikanth S Narayanan, et al., "The interspeech 2010 paralinguistic challenge.," in *Interspeech*, 2010, vol. 2010, pp. 2795–2798.

[2] Björn W Schuller, Stefan Steidl, Anton Batliner, Simone Hantke, Florian Hönig, Juan R Orozco-Arroyave, Elmar Nöth, Yue Zhang, and Felix Weninger, "The interspeech 2015 computational paralinguistics challenge: nativeness, parkinson's & eating condition.," in *INTERSPEECH*, 2015, pp. 478–482.

[3] Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proceedings of Interspeech*, 2016.

[4] Bj"orn Schuller, Stefan Steidl, Anton Batliner, Elika Bergelson, Jarek Krajewski, Christoph Janott, Andrei Amatuni, Marisa Casillas, Amanda Seidl, Melanie Soderstrom, et al., "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring," .

[5] Hung-Shin Lee, Yu-Ding Lu, Chin-Cheng Hsu, Yu Tsao, Hsin-Min Wang, and Shyh-Kang Jeng, "Discriminative autoencoders for speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5375–5379.

[6] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, MIT Press, 2016.

[7] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori, "Speech enhancement based on deep denoising autoencoder.," in *Interspeech*, 2013, pp. 436–440.

[8] Xue Feng, Yaodong Zhang, and James Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1759–1763.

[9] Wei-Jen Lee, Syu-Siang Wang, Fei Chen, Xugang Lu, Shao-Yi Chien, and Yu Tsao, "Speech dereverberation based on integrated deep and ensemble learning," *arXiv preprint arXiv:1801.04052*, 2018.

[10] Takaaki Ishii, Hiroki Komiyama, Takahiro Shinozaki, Yasuo Horiuchi, and Shingo Kuroiwa, "Reverberant speech recognition based on denoising autoencoder.," in *Interspeech*, 2013, pp. 3512–3516.

[11] N. Zhou and J. Fan, "Jointly learning visually correlated dictionaries for large-scale visual recognition applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 4, pp. 715–730, April 2014.

[12] Fabien Scalzo, George Bebis, Mircea Nicolescu, Leandro Loss, and Alireza Tavakkoli, "Feature fusion hierarchies for gender classification," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.

[13] Jing Han, Zixing Zhang, Nicholas Cummins, Fabien Ringeval, and Bjrn Schuller, "Strength modelling for real-worldautomatic continuous affect recognition from audiovisual signals," *Image and Vision Computing*, pp. –, 2016.

[14] Manhua Liu, Daoqiang Zhang, and Dinggang Shen, "Hierarchical fusion of features and classifier decisions for alzheimer's disease diagnosis," *Human brain mapping*, vol. 35, no. 4, pp. 1305–1319, 2014.

[15] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[16] Heysem Kaya and Alexey A Karpov, "Fusing acoustic feature representations for computational paralinguistics tasks," *Interspeech 2016*, pp. 2046–2050, 2016.

[17] Fréjus AA Laleye, Eugène C Ezin, and Cina Motamed, "Speech phoneme classification by intelligent decision-level fusion," in *Informatics in Control, Automation and Robotics 12th International Conference, ICINCO 2015 Colmar, France, July 21-23, 2015 Revised Selected Papers*. Springer, 2016, pp. 63–78.

[18] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, New York, NY, USA, 2013, MM '13, pp. 835–838, ACM.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[20] Brian McFee, Matt McVicar, Oriol Nieto, Stefan Balke, Carl Thome, Dawen Liang, Eric Battenberg, Josh Moore, Rachel Bittner, Ryuichi Yamamoto, Dan Ellis, Fabian-Robert Stoter, Douglas Repetto, Simon Waloschek, CJ Carr, Seth Kranzler, Keunwoo Choi, Petr Viktorin, Joao Felipe Santos, Adrian Holovaty, Waldir Pimenta, and Hojin Lee, "librosa 0.5.0," Feb. 2017.

[21] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Dec. 2011, IEEE Signal Processing Society, IEEE Catalog No.: CFP11SRW-USB.

[22] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks.," in *Aistats*, 2010, vol. 9, pp. 249–256.

[23] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.