BRIDGING MIXTURE DENSITY NETWORKS WITH META-LEARNING FOR AUTOMATIC SPEAKER IDENTIFICATION

Ruirui Li[⋆] *Jyun-Yu Jiang*[⋆] *Xian Wu*[¶] *Hongda Mao*[§] *Chu-Cheng Hsieh*[†] *Wei Wang*[⋆]

[⋆]University of California Los Angeles [¶]University of Notre Dame [§†] Amazon
[⋆]{rrli, jyunyu, weiwang}@cs.ucla.edu, [¶]xwu9@nd.edu, [§]hdmao@mail.rit.edu, [†]chucheng@ucla.edu

ABSTRACT

Speaker identification answers the fundamental question "Who is speaking?" The identification technology enables various downstream applications to provide a personalized experience. Both the prevalent i-vector based solutions and the state-of-the-art deep learning solutions usually treat all users equally, with no distinctions between new users and existing users, during the training process. We notice that a good many new users start with limited labeled training data, which often results in inferior predicting performance of recognizing users' voices. To alleviate the disadvantage caused by training data deficiency, we propose a Mixture Density Network-based Meta-Learning method (*MDNML*) for speaker identification. *MDNML* emphasizes the expeditious process of learning to recognize new users where each has only a few seconds of labeled data.

We conduct experiments on the LibriSpeech dataset and compare *MDNML* with four state-of-the-art baseline methods. The results conclude that *MDNML* achieves higher accuracy in recognizing new users with limited labeled utterances than all baseline methods. Our proposed solution significantly expedites the learning by transferring the knowledge learned from the existing user base through gradient-based meta-learning. We consider our work to be a stepping-stone for more sophisticated meta-learning frameworks for accelerating voice recognition. Furthermore, we discuss a strategy for enhancing the accuracy by incorporating the notion of household-based acoustic profiles with *MDNML*.

***Index Terms*—** mixture density networks, meta-learning, new users, speaker identification

1. INTRODUCTION

A recent report[1] in 2018 shows that smart speakers have gained an installed user base of nearly one in every four U.S. adults or 50+ million users. These smart speakers equipped with voice recognition technology, also known as speaker identification, which answers the fundamental question "Who is speaking?" The answer to the question enables various downstream applications to provide a personalized experience.

Speaker identification has been studied as a supervised classification problem based on the characteristics of voices. Traditionally, the prevalent solution is based on i-vector representation of speech segments [1], which is combined with the improvement over the Gaussian Mixture Model-Universal Background Models (GMM-UBMs) [2]. Recently, deep learning-based methods have been gaining attraction as they outperformed the prevalent i-vector solutions. Deep speaker [3] utilizes residual networks to extract audio embeddings and optimizes triplet loss to perform speaker identification. VGGVox [4] adopts a CNN-based residual network to construct audio embeddings and optimizes contrastive loss with pre-training based on softmax classification. Two other Resnet-based methods utilize additive margin softmax classification loss [5] to improve the recognition accuracy in [6, 7]. SincNet [8] utilizes convolutional neural networks to learn speaker recognition directly from raw audios.

Our work presumes new users always have very limited labeled voice data, as Google Assistant and Amazon Alexa only require a new user to repeat two to four prompts for learning his/her voice. Unlike the aforementioned research where existing users and new users are treated equally, we develop a meta-learning approach targeting to expedite the learning process for recognizing new users with limited training data. We foresee the need of expediting the learning for new users as (1) smart speakers are gaining in popularity where the report also shows that 30% of users are new in 2018 and (2) the previous research [9] displayed that the length of voice history of a user is positively correlated to his/her identification accuracy.

Our proposed solution expedites the learning by transferring the knowledge learned from the existing user base with a gradient-based meta-learning tactic (Section 2.2). We use Mixture Density Networks (MDNs) [10] (Section 2.1) to construct acoustic user profiles in that MDNs are gradient-friendly and can model voice utterances with arbitrary lengths so that we can then apply Model-Agnostic Meta-Learning (MAML) [11] technique to achieve expeditious learning. Our experiments demonstrate that our proposed solution, *MDNML*, when having only four seconds of voice data from

[1]https://voicebot.ai/wp-content/uploads/2018/11/voice-assistant-consumer-adoption-report-2018-voicebot.pdf

new users, its accuracy outperforms the best/worst baseline methods by 3.2%/5.8% (Section 3.2).

2. PROBLEM STATEMENT & METHODOLOGY

We formulate the objective of our work as the following. Suppose the system has a set of existing users with registered voice utterances as background training data. Given a set of new users, with a short registered voice utterance for each user as enrollment, and another short testing voice utterance of a user within the new user set, the goal of this study is to recognize the speaker identity behind the testing voice utterance. For simplicity, we compare model performance based on text-independent tasks and presume that new users have very limited training data [9, 12–14], for example, one to four seconds.

To better explain how to construct users' acoustic profiles and how to transfer profiling knowledge from existing users to new users, we illustrate the framework of *MDNML* in Fig. 1.

2.1. Mixture Density Networks

Mixture density networks (MDNs) are based on a mixture density model that combines neural networks [10]. MDNs are chosen in this work to construct acoustic profiles for users as they are inherently flexible in sense that it can model voice utterances with arbitrary lengths. Moreover, assuming the voice print of a user can be sufficiently expressed by a short period of time, each tiny time frame can contribute to one training instance for the user, leading to a relatively adequate amount of training data for new users. In addition, MDNs based on neural networks are gradient-friendly so that the gradient-based knowledge transfer techniques are applicable.

In this work, we utilize mel-frequency cepstral coefficients (MFCCs) [15] to represent the voice characteristics of users because MFCCs are capable of approximating the human aural systems and widely applied in various voice recognition tasks, such as speaker recognition [13, 16–18] and speech synthesis [19–22]. More specifically, we utilize a Gaussian mixture model (GMM)-based MDN. An MDN maps a set of input MFCC features x to the parameters of a GMM (i.e., mixture weights π_m, mean μ_m, and variance σ_m^2), which in turn give a full probability density function of a MFCC feature y, conditioned on the input x and the learned model $\mathcal{M}$, $p(y \mid x, \mathcal{M})$, Formally,

$$p(y \mid x, \mathcal{M}) = \sum_{m=1}^{M} \pi_m(x) \cdot \Phi(y; \mu_m(x), \sigma_m^2(x)), \tag{1}$$

where M is the number of mixture components and $\pi_m(x)$, $\mu_m(x)$, and $\sigma_m^2(x)$ correspond to the mixture weight, mean, and variance of the m-th component conditioned on x. Φ is the Gaussian mixture component.

To derive the parameters in a GMM-based MDN, MDN first converts the input x using a multi-layer perceptron (MLP) and obtains output z as:

$$z = f_\theta(x), \tag{2}$$

where $f_\theta(\cdot)$ corresponds to a set of transformations in the MLP network and θ denotes the set of parameters to be learned. The total number of network outputs, i.e., the dimension of z, is $(2c + 1) \times M$ where c corresponds to the dimension of the MFCC features. M corresponds to the number of mixture components in the MDN. Then, z is partitioned into three subsets $z_m^{(\pi)} \in \mathbb{R}^{1\times 1}$, $z_m^{(\mu)} \in \mathbb{R}^{1\times c}$, and $z_m^{(\sigma)} \in \mathbb{R}^{1\times c}$, which correspond to the outputs used to calculate the GMM weights, means, and standard derivations, respectively.

$$z = [z_1^{(\pi)}, ..., z_M^{(\pi)}, z_1^{(\mu)}, ..., z_M^{(\mu)}, z_1^{(\sigma)}, ..., z_M^{(\sigma)}]. \tag{3}$$

After the partition, each subset is passed through a set of specific transformations for conversion to the GMM weights, means, and standard derivations as:

$$\pi_m(x) = \frac{\exp(z_m^{(\pi)})}{\sum_{j=1}^{M} \exp(z_j^{(\pi)})}, \tag{4}$$

$$\mu_m(x) = \tanh(z_m^{(\mu)}), \tag{5}$$

$$\sigma_m(x) = \exp(z_m^{(\sigma)}). \tag{6}$$

The use of the softmax function in Equation 4 constrains the mixture weights to be positive and sum up to 1. Analogously, Equation 6 constrains the standard deviations to be positive.

During training, these density parameters are passed to a log likelihood calculator to compute the log likelihood of an MFCC feature y, which is further utilized to define the loss function for the MDN as follows:

$$L = -\sum_{n=1}^{N} \log\{\sum_{m=1}^{M} \pi_m(x) \cdot \Phi(y; \mu_m(x), \sigma_m^2(x))\}, \tag{7}$$

where N is the number of MFCC vectors for a user. The parameters of MDN only lie in the MLP network and these parameters are optimized in such a way that the overall negative log likelihood in Equation 7 is minimized.

2.2. Knowledge Transfer via Gradient-based Meta-learning

The effective training of MDNs relies on sufficient training data, which are usually unavailable for new users. To compensate for the data deficiency, we develop a gradient-based knowledge transfer module to leverage identification knowledge grained from recognizing existing users. More precisely, we learn a set of well-initialized model parameters over many similar tasks so that it would be easier to reach the global optimal when training a new task.

Each task corresponds to the training process of creating an acoustic profile of a user, where a profile is expressed by an MDN. We optimize a set of parameters Ψ such that when a gradient step is taken with respect to particular task t_i, the

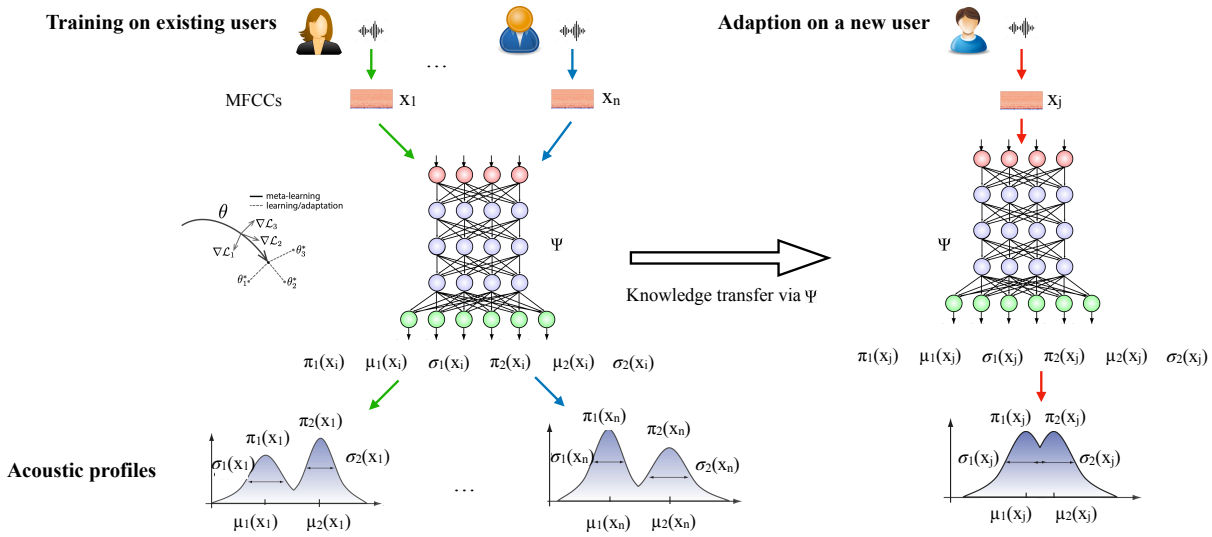

Fig. 1: The framework of *MDNML*. During training, we learn a set of well-initialized model parameters Ψ by training acoustic profiles of all existing users. To serve new users, we construct their acoustic profiles by adapting from Ψ.

parameters θ_i, derived from Ψ, are close to the optimal parameters for task t_i, where $\theta_i = \{\pi, \mu, \sigma\}$ denotes the model parameters learned based on task t_i. Let $l(\theta_i)$ denote the loss of task t_i based on the test set of t_i. The entire loss over multiple tasks is given by:

$$L(\Psi) = \sum_{i=1} l(\theta_i). \tag{8}$$

To update the initialization parameters Ψ, we have:

$$\Psi \leftarrow \Psi - \alpha \nabla_{\Psi} L(\Psi). \tag{9}$$

To optimize each individual task t_i, we have:

$$\theta_i \leftarrow \Psi - \beta \nabla_{\theta_i} l(\Psi), \tag{10}$$

where α is the meta-learning rate, and β is the learning rate for each individual task, i.e., the training of a mixture density model. Algorithm 1 shows the detailed training and adaption processes of *MDNML*.

Algorithm 1: Acoustic profile training and adaption

Input: learning rate α, meta-learning rate β, maximal number of iterations itr_{max}, inner update size T_{train} in training, inner update size T_{adapt} in adaption

/* Training on the existing users */
for $itr \leq itr_{max}$ **do**
  Sample a batch of existing users as U
  for user i in U **do**
    Sample a piece of audio of user i
    $\theta_i^{(0)} = \Psi$
    for $t \leq T_{train}$ **do**
      $\theta_i^{(t)} = \theta_i^{(t-1)} - \alpha \nabla_{\theta_i^{(t-1)}} L(\theta_i^{(t-1)})$
  $\Psi = \Psi - \beta \nabla_{\Psi} \sum_{i \in U} L(\theta_i^{(T)})$

/* Adaption on new users by fine-tuning*/
for new user j in U_{adapt} **do**
  $\theta_j^{(0)} = \Psi$
  for $t \leq T_{adpat}$ **do**
    $\theta_j^{(t)} = \theta_j^{(t-1)} - \alpha \nabla_{\theta_j^{(t-1)}} L(\theta_j^{(t-1)})$

2.3. Speaker Identification in a Household

We now discuss how to utilize the constructed users' acoustic profiles to conduct speaker identification given a short voice utterance of a user in a household. Following GMM-UBM [2], in addition to training an acoustic profile $\mathcal{M}_i$ for each user i in the household, we also train a household-level background acoustic profile $\mathcal{M}_{hbm}$ using the mixtures of all training utterances of the users in the household.

Given a user's short voice utterance x_j, we feed it into the universal background profile and each individual acoustic profile, with each profile yielding a vector of fitness scores. Each vector of scores indicates how well the voice utterance fit the corresponding acoustic profile. More specifically, we use $p(x_j \mid \mathcal{M}_{hbm})$ and $p(x_j \mid \mathcal{M}_i)$ to denote the scores for the household-level profile and the profile of user i in the household, respectively. Formally, the speaker identify is given by:

$$\arg\max_i f(\mathbf{1}_{>0}(p(x_j \mid \mathcal{M}_i) - p(x_j \mid \mathcal{M}_{hbm}))), \tag{11}$$

where $\mathbf{1}_{>0}(\cdot)$ is the vector-level indicator function and $f(\cdot)$ is a counter, which calculates the number of 1's in its input. By introducing the household-level background profile, it allows us to achieve speaker identification based on background-proof voice frames, which potentially offers stronger discriminative power.

3. EXPERIMENTS

We conduct experiments on LibriSpeech data set to evaluate the performance of *MDNML* against four popular algorithms.

3.1. Dataset and Experimental Settings

The experiments are conducted on the LibriSpeech dataset, which is publicly available[2]. Table 1 shows the number of speakers in the dataset. For the dataset, 75% of speakers are treated as existing users and the remaining 25% of speakers are treated as new users for the purpose of evaluation. We follow the previous work [14] to extract acoustic features from the raw audios. The first 20 MFCCs are extracted from speech after an energy-based voice activity detection. A 44 kHz sampling rate and a 25 ms hamming window with a 10 ms frame shift are used during the MFCC construction.

Table 1: The number of speakers in the experimental dataset.

Dataset	#(Female speakers)	#(Male speakers)	#(Total speakers)
LibriSpeech	125	126	251

Baseline Methods. To evaluate the performance of *MDNML*, the following four methods are adopted as baselines.

- **MDN** [10] trains acoustic profiles for each new user from scratch without any knowledge gained from existing users.
- **PN** [14] utilizes the CNN-based prototypical network, a metric-learning-based few-shot technique, to conduct speaker identification.
- **PNL** [13] relies on Bi-LSTM-based prototypical network to perform speaker identification.
- **AFEASI** [9] applies adversarial training on prototypical network to achieve speaker identification.

3.2. Identification Performance

In this section, we evaluate the performances of *MDNML* against different baseline methods on the LibriSpeech dataset. We adopt household-level accuracy as the evaluation metric. The household-level accuracy first calculates the identification accuracy in each household and then averages the identification accuracy in each household by treating the importance of them equally.

To imitate the scenarios of serving new users, we set the duration of each enrollment utterance, which is used for profile adaptions for the new users, to small values, varying from 2 to 4 seconds. Moreover, in order to offer instant identification response, we vary the duration of each test voice utterance from 1 second to 4 seconds. Tables 2 and 3 show the performance of different methods on the LibriSpeech dataset.

We have four observations from the results on the LibriSpeech dataset. First, in general, the longer the duration of a voice utterance for identification in test, the higher accuracy each method can achieve. For example, when we have 2 seconds long voice utterance as the enrollment for profile adaption and 1 second long voice utterance to identify the speakers with testing data, MDN can reach only 80% accuracy; however, the accuracy increases to 83.4%, 86.2%, and 87.8% when the test voice utterance becomes to 2, 3, and 4 seconds, respectively. Note that this observation generally applies to all methods. The performance gain stems from the fact that the acoustic signals embedded in long voice utterances are more consistent and reliable. These consistent and reliable acoustic signals further yield confident speaker identifications, which are more accurate. Second, the longer the voice enrollment we have for a new user during the adaption process, the higher accuracy each method can achieve. It makes sense because the longer the voice enrollment for a user, the richer signals we can use to construct his/her acoustic profile by analyzing the enrolled utterance. Profile constructions, supported by rich acoustic information, contribute to the high accuracy. Third, we observe that PN, PNL, AFEASI, and *MDNML* generally outperform MDN. This shows the advantages of utilizing few-shot learning, which allows effective learning even with limited data. Fourth, *MDNML* consistently achieves the highest accuracy comparing with the four baselines in all settings. It demonstrates the effectiveness of *MDNML*, which leverages knowledge learnt from existing users.

Table 2: Accuracy with 2s voice enrollment.

Utterance duration in test	1s	2s	3s	4s
MDN	80.0%	83.4%	86.2%	87.8%
PN	85.8%	86.0%	88.0%	89.4%
PNL	82.8%	86.4%	85.8%	85.0%
AFEASI	86.6%	86.6%	89.4%	90.2%
MDNML	88.6%	88.6%	90.2%	91.4%

Table 3: Accuracy with 4s voice enrollment.

Utterance duration in test	1s	2s	3s	4s
MDN	85.8%	86.4%	88.2%	89.0%
PN	88.0%	87.6%	88.8%	89.2%
PNL	84.8%	86.0%	86.8%	87.2%
AFEASI	88.6%	89.0%	89.2%	89.8%
MDNML	89.6%	90.4%	92.2%	93.0%

4. CONCLUSION

We present *MDNML*, a framework that combines mixture density network (in acoustic profile creation) and model-agnostic meta-learning (in transferring knowledge) to achieve expeditious learning for speaker identification tasks. The proposed solution alleviates the unpleasant difficulty of learning (new) user voice with training data deficiency. Our experiments on LibriSpeech dataset show that the proposed solution outperforms the four baseline methods: MDN, PN, PNL, and AFEASI.

[2]LibriSpeech: http://www.openslr.org/12

5. REFERENCES

[1] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Trans. Audio, Speech & Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[2] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[3] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu, "Deep Speaker: an End-to-End Neural Speaker Embedding System," *CoRR*, vol. abs/1705.02304, 2017.

[4] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep Speaker Recognition," *CoRR*, vol. abs/1806.05622, 2018.

[5] Feng Wang, Weiyang Liu, Hanjun Dai, Haijun Liu, and Jian Cheng, "Additive Margin Softmax for Face Verification," in *Proceedings of ICLR, Vancouver, BC, Canada, April 30 - May 3*, 2018.

[6] Weidi Xie, Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Utterance-level Aggregation for Speaker Recognition in the Wild," in *Proceedings of ICASSP, Brighton, United Kingdom, May 12-17*, 2019, pp. 5791–5795.

[7] Mahdi Hajibabaei and Dengxin Dai, "Unified Hypersphere Embedding for Speaker Recognition," *CoRR*, vol. abs/1807.08312, 2018.

[8] Mirco Ravanelli and Yoshua Bengio, "Speaker Recognition from Raw Waveform with SincNet," in *2018 IEEE Spoken Language Technology Workshop, SLT, Athens, Greece, December 18-21*, 2018, pp. 1021–1028.

[9] Ruirui Li, Jyun-Yu Jiang, Jiahao Liu Li, Chu-Cheng Hsieh, and Wei Wang, "Automatic speaker recognition with limited data," in *WSDM, Houston, TX*, 2020, pp. 340–348.

[10] Christopher M Bishop, "Mixture Density Networks," 1994.

[11] Chelsea Finn, Pieter Abbeel, and Sergey Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," in *Proceedings of ICML, Sydney, NSW, Australia, 6-11 August*, 2017, pp. 1126–1135.

[12] L. Wan, Q. Wang, A. Papir, and I. Moreno, "Generalized End-to-end Loss For Speaker Verification," in *Proceedings of ICASSP, Calgary, AB, Canada, April 15-20*, 2018, pp. 4879–4883.

[13] J. Wang, K. Wang, M. Law, F. Rudzicz, and M. Brudno, "Centroid-based Deep Metric Learning For Speaker Recognition," in *Proceedings of ICASSP, Brighton, United Kingdom, May 12-17*, 2019, pp. 3652–3656.

[14] Prashant Anand, Ajeet Kumar Singh, Siddharth Srivastava, and Brejesh Lall, "Few Shot Speaker Recognition using Deep Neural Networks," *CoRR*, vol. abs/1904.08775, 2019.

[15] Min Xu, Ling-Yu Duan, Jianfei Cai, Liang-Tien Chia, Changsheng Xu, and Qi Tian, "HMM-Based Audio Keyword Generation," in *Proceedings of PCM, Tokyo, Japan, November 30 - December 3*, 2004, pp. 566–574.

[16] Zheli Liu, Zhendong Wu, Tong Li, Jin Li, and Chao Shen, "GMM and CNN hybrid method for short utterance speaker recognition," *IEEE Transactions on Industrial informatics*, vol. 14, no. 7, pp. 3244–3252, 2018.

[17] Ankur Maurya, Divya Kumar, and RK Agarwal, "Speaker recognition for Hindi speech signal using MFCC-GMM approach," *Procedia Computer Science*, vol. 125, pp. 880–887, 2018.

[18] Ondřej Novotnỳ, Oldřich Plchot, Ondřej Glembek, Lukáš Burget, and Pavel Matějka, "Discriminatively re-trained i-vector extractor for speaker recognition," in *Proceedings of ICASSP, Brighton, United Kingdom, May 12-17*, 2019, pp. 6031–6035.

[19] Lauri Juvela, Bajibabu Bollepalli, Xin Wang, Hirokazu Kameoka, Manu Airaksinen, Junichi Yamagishi, and Paavo Alku, "Speech Waveform Synthesis from MFCC Sequences with Generative Adversarial Networks," in *Proceedings of ICASSP, Calgary, AB, Canada, April 15-20, 2018*, 2018, pp. 5679–5683.

[20] Ruibo Fu, Jianhua Tao, Zhengqi Wen, and Yibin Zheng, "Phoneme dependent speaker embedding and model factorization for multi-speaker speech synthesis and adaptation," in *Proceedings of ICASSP, Brighton, United Kingdom, May 12-17*, 2019, pp. 6930–6934.

[21] Lauri Juvela, Bajibabu Bollepalli, Xin Wang, Hirokazu Kameoka, Manu Airaksinen, Junichi Yamagishi, and Paavo Alku, "Speech waveform synthesis from mfcc sequences with generative adversarial networks," in *Proceedings of ICASSP, Calgary, AB, Canada, April 15-20*, 2018, pp. 5679–5683.

[22] Kévin Vythelingum, Yannick Estève, and Olivier Rosec, "Acoustic-dependent phonemic transcription for text-to-speech synthesis.," in *Interspeech*, 2018, pp. 2489–2493.