Vol.2018-MUS-118 No.22
Vol.2018-SLP-120 No.22
2018/2/21

劣化音声を用いたDNN音声合成のための話者類似度に基づく教師なし話者適応

高木 信二[1,a)]　西村 祥一郎[2,b)]　山岸 順一[1,c)]

概要：我々はこれまで，DNN に基づく音声合成において，話者適応にテキストを必要としない，教師なし話者適応を提案した．この手法は，話者認識において広く用いられているモデル (GMM-UBM，i-vector/PLDA) を利用し計算された，個々の学習話者に対する事後確率を連結したベクトルにより話者類似度が表現されると仮定し，目標話者の音声から計算された話者類似度ベクトルを，DNN 音声合成システムの入力コードとして用いることで実現される．本論文では，目標話者の音声が雑音や残響を含み劣化していることを想定し，このような劣化音声に対し，ロバストな教師なし話者適応について検討する．具体的には，雑音や残響を付与した音声データを用いることで，目標話者の劣化音声からであっても，適切に話者類似度が可能な話者認識モデルの構築を行う．10 代後半から 80 代までの話者がバランス良く含まれた 135 名からなる巨大コーパスを用い，評価実験を行った．客観評価の結果より，適切に話者認識モデルを構築することで，劣化音声にロバストな教師なし話者適応が可能であることが確認できた．

キーワード：音声合成，DNN，教師なし話者適応，話者認識

1. はじめに

高い柔軟性を持つ DNN に基づく音声合成システムの研究は広く行われており，DNN に基づく音響モデルにおいて，複数話者モデリングや様々な話者適応手法が提案されている．例えば，話者情報を DNN の入力とすることで複数話者の音声の合成を行う手法 [1], [2] が提案されており，また，話者適応には話者情報の推定 [3]，i-vector[4]，GMM を用いた出力の変換 [4]，Hidden UnitContribution に基づく適応 (LHUC)[5] 等が利用されている．その他，複数話者データを用いることで，隠れ層は全話者で共有されるが話者特有の出力層を持つ音響モデルの構築が行なわれている [6]．この手法では回帰層のみを推定することによる話者適応が検討されている．

我々はこれまで，言語特徴量に加え追加の特徴量として話者・ジェンダー・年齢コード (入力コードと呼ぶ) を利用した DNN 音声合成に基づく，高精度な複数話者モデリング，及び，話者適応を提案した [7]．また，この手法を拡張し，目標話者の音声データのみを必要とし，テキストを必要としない教師なし話者適応を提案した [8]．この教師なし話者適応は，入力コードに基づく DNN 音声合成システムにおいて，音声のみから計算可能な入力コードを話者コードとして利用し，話者適応時にも未知話者の話者コードを音声のみから推定することにより実現される．入力コードとして，話者認識において広く用いられているモデル (GMM-UBM，i-vector/PLDA) を利用し計算された，学習話者に対する事後確率を利用する．ここでは，各学習話者に対する事後確率が，学習話者の違いの類似度を表現していると仮定し，DNN 音声合成のための複数話者モデルの構築を行う．学習された DNN 音声合成モデルは，学習話者の違いの類似度を反映し，さらに，入力される話者類似度ベクトルが変化した場合には，出力音声も変化すると期待できる．これにより，合成音声の品質を下げることなく，高精度な教師なし話者適応が可能であることを確認した．

しかし，[8] では雑音や残響が含まれない，高品質音声データが目標話者の音声データとして用いられいる．このような高品質音声データの収録は高価な機材，防音室等の収録環境を準備する必要があり高コストである．本論文では，目標話者の音声が雑音や残響により劣化していること

[1] 国立情報学研究所
National Institute of Informatics, Chiyoda, Tokyo 101–8430, Japan
[2] 株式会社オルツ
alt Inc., The Canal Gate Akihabara 8F, 3-1-2 Higashikanda, Chiyoda-ku, Tokyo, Japan
[a)] takaki@nii.ac.jp
[b)] yoshikazu.nishimura@alt.ai
[c)] jyamagis@nii.ac.jp

を想定し，このような劣化音声に対し，ロバストな教師なし話者適応について検討する．本研究で用いる DNN に基づく音声合成では，話者認識モデルを利用し計算された，学習話者に対する事後確率が入力コードとして利用される．劣化音声を用いた場合でも，学習話者に対する事後確率を適切に計算可能であれば，高品質な教師なし話者適応が期待できる．本論文では，劣化音声を学習データとして利用することで，劣化音声に対しロバストな話者認識モデルの構築を検討し，教師なし話者適応の評価を行う．

2. 入力コードを用いた DNN 複数話者音声合成モデル

ここでは，文献 [7] の手法を簡潔にレビューする．DNN に基づく複数話者音声合成システムは，複数話者コーパスを用い特定話者モデル構築と同様の手順で構築を行う．しかし，各話者の音響特性を保持することに加え話者適応を可能とするため，言語特徴量に加え追加の特徴量 (入力コードと呼ぶ) を DNN 音声合成システムの入力として用いる．これら入力コードを用いることで，話者，ジェンダー，年齢等を区別した DNN に基づく音響モデルの学習，また，音声合成を行うことが期待できる．

入力コードとして話者・ジェンダー・年齢コードを用いる．例えば，話者コードに One-hot ベクトル，ジェンダーコードにバイナリ値 (0 が女性，1 が男性)，年齢コードに年齢そのものを表現する 1 次元の数値が利用される．

3. 話者類似度に基づく教師なし話者適応

本研究では，入力コードに基づく DNN 音声合成における，テキストを必要としない教師なし話者適応 [8] を利用する．本手法における，音声合成のための複数話者モデルの構築，話者適応の手順は以下の通りである．

(1) 個々の学習話者に対する話者認識モデルを構築する．利用する音声データは DNN に基づく複数話者音声合成システム構築に用いる音声データと同様である．本研究では，話者認識モデルに GMM-UBM[9]，または，i-vector/PLDA[10] を利用した．
(2) Step.1 で構築した話者認識モデルと学習話者の音声データを用い，個々の学習話者に対する事後確率を計算する．得られた事後確率を連結し，話者類似度ベクトルとする．本研究では，学習データには学習話者が 112 人含まれるため，112 次元のベクトルにより話者類似度が表現される．
(3) Step.2 で求めた話者類似度ベクトルを話者コードとし，ジェンダー・年齢コードを加えた入力コードを用い，DNN に基づく複数話者音声合成システムの構築を行う．
(4) Step.1 で構築した話者認識モデルを用い，目標話者の適応データから話者コードを表現する話者類似度ベクトルを推定する

これにより，話者類似度ベクトルに基づく DNN に基づく複数話者音声合成システムの構築，目標話者に対する話者類似度ベクトルの推定が行われ，教師なし話者適応が可能となる．

4. 話者認識モデル構築に用いる学習データ

入力コードに基づく DNN 音声合成における教師なし話者適応では，話者認識モデルは個々の学習話者に対する事後確率を求めるために用いられる．また，得られた個々の学習話者に対する事後確率を連結したベクトルを話者コードとし，DNN 音声合成の入力として用いる．そのため，話者認識モデルの性能が出力される合成音声に大きな影響を与える．

[8] では，セクション 3 に記したように，話者認識モデルの学習に，音声合成システム構築に用いる劣化のない高品質音声データが用いられている．本研究では，この高品質音声データに雑音，残響を付与し劣化させ，話者認識モデルの学習を行うことで，劣化音声にロバストな話者認識モデルの構築を検討する[*1]．劣化音声の作成には，雑音付与のため Demand データベース [11]，残響付与のため the ACE Challenge データベース [12] を用いた．オフィスルーム，ミーティングルームでの音声収録を想定し，Demand データベースからは OFFICE，MEETING の 48kHz サンプリング，チャネル 1 の 2 種類の雑音データを，the ACE Challenge データベースからは OFFICE1，MEETING ROOM1 の 2 種類の部屋のインパルス応答を選択した．文献 [13] と同様に，以下の通り劣化音声 y を作成した．

$$y = x * h_1 + \alpha(n * h_2). \quad (1)$$

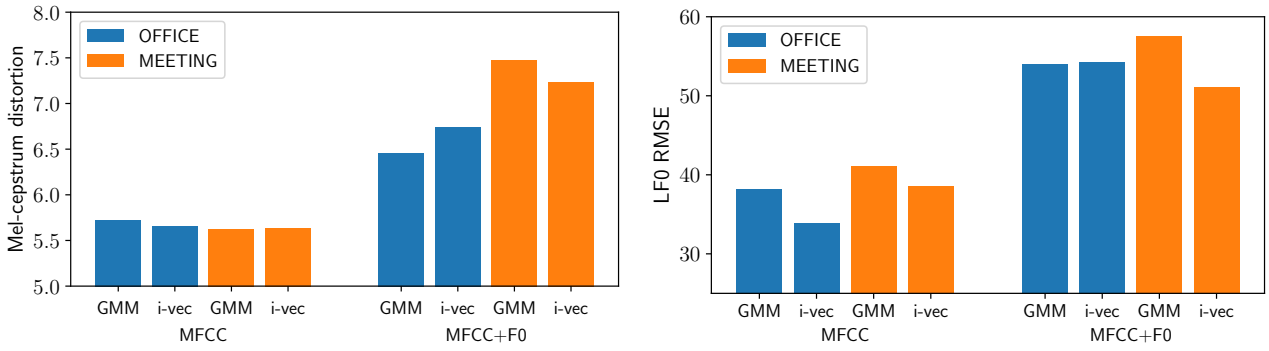
ここで，x と n はそれぞれ高品質音声と雑音を，h_1 と h_2 はそれぞれ異なるマイク位置 (h_1 は h_2 よりスピーカに近い) で得られたルームインパルス応答を，$*$ は畳み込みを表す．また，α は雑音の強さを調整するパラメータであり，所望の SNR の劣化音声を作成するのに用いる．また，実験においては話者認識モデルの学習データだけでなく，目標話者の音声データについても同様の手法で劣化させた．

5. 実験

5.1 実験条件

実験には，10 代後半から 80 代までの男性 65 名，女性 70 名からなる高品質日本語コーパスを用いた．このコーパスから男性 56 名，女性 56 名の音声データを複数話者モデルの学習に用い，残りの男性 9 名，女性 14 名の音声データ

[*1] DNN に基づく複数話者音声合成システムの構築に用いる個々の学習話者に対する事後確率は，劣化のない高品質音声データから計算した．



(a) メルケプストラム歪み　(b) LF0 RMSE

図 1 異なる音響特徴量 (MFCC, MFCC+F0) と手法 (GMM, i-vec) を用いて構築した話者認識モデルにおける，教師なし話者適応の結果 (メルケプストラム歪み，LF0 RMSE)．ここで GMM，i-vec はそれぞれ GMM-UBM と i-vector/PLDA を表す．また，ラベル (OFFICE，MEETING) は話者認識モデルの学習，適応データとして用いた劣化音声の種類を表す．

を適応実験に用いた．学習セットに含まれる話者は年齢，ジェンダーがバランス良く含まれるように設定し，各年齢帯とジェンダーそれぞれで 8 名である．発話数も話者間で大きな偏りがないように設定し各話者それぞれ約 100 文を用い，合計で 11,154 発話である．

適応実験では，学習データに含まれない 23 話者から 100 文を適応データとして用いた．実験では，適応データを劣化させずそのまま高品質音声データ (CLEAN) として利用するか，オフィスルームの雑音・残響の付与 (OFFICE)，ミーティングルームの雑音・残響の付与 (MEETING) を行った．様々な強度の劣化の影響を見るため，SNR が 2.5dB，7.5dB，12.5dB，17.5dB の劣化音声を作成した．また，評価実験では，適応データの劣化の種類 (OFFICE，MEETING) は既知としたが，劣化の強度は未知とした．

話者認識モデルの学習には，SIDEKIT[14] を用いた．特徴量には 0 次を含む 19 次 MFCC とその Δ，Δ^2 も用いた．また，F0 に関する特徴量の利用も検討した．F0 に関する特徴量としては，F0 をそのまま用いるとスペクトル特徴量と次元数が大きく異なることから，当該フレームと前後 32 フレームの無声部を補完した log F0 に対して DCT を行い 20 次元にした特徴量とその Δ，Δ^2(F0) を利用した．これら，特徴量を用い GMM-UBM の学習，i-vector 抽出のための UBM の学習を行なった．GMM のミクスチャ数は 64，i-vector の次元数は 400，G-PLDA の話者部分空間に対する基底数は 20 である．

劣化なし高品質音声データ (CLEAN)，オフィスルームの雑音・残響を付与した劣化音声 (OFFICE)，ミーティングルームの雑音・残響を付与した劣化音声 (MEETING) の 3 種類の音声を用い，それぞれ話者認識モデルを構築した．話者認識モデルの学習では，様々な強度の劣化に対応するため，劣化音声を SNR を 0.0dB，5.0dB，10.0dB，15.0dB からランダムに選択し作成した．3 種類の音声について，異なる音響特徴量 (MFCC, MFCC+F0)，モデル化手法 (GMM-UBM, i-vector/PLDA) を用い，合計 12 種類の話者認識モデルを構築した．

DNN 音声合成システムの構築には，0 次を含む 59 次メルケプストラム係数，無声部を補完した log F0，無声/有声パラメータ，25 次非周期成分を用いた．これら音響特徴量は F0 適応窓を用いて WORLD により各フレーム毎に抽出を行った．フレームシフトは 5ms である．また，無声/有声パラメータ以外のパラメータについてはその Δ，Δ^2 も用いた．DNN に基づく音響モデル構築に利用した音響特徴量は合計 259 次元となる．

言語特徴量は 389 次元であり，DNN に基づく音響モデルの入力として用いられる．言語特徴量に含まれる音素継続長は HMM により学習データ，テストデータともに Forced-alignment により得た．つまり，音声合成時にテストデータから得られた音素継続長を利用している．また，実験では言語特徴量に加え話者，ジェンダー，年齢を表現する入力コードを DNN に基づく音響モデルの入力として利用する．

音響モデルは隠れ層数 5 で各隠れ層が 1024 ユニットからなるフィードフォワード DNN を用いた．シグモイド関数を隠れ層，出力層の活性化関数として用いた．音響モデルのパラメータはランダム値により初期化を行い，学習率 (0.05)，学習回数 (10 epochs)，ミニバッチサイズ (256 フレーム) を固定し確率的勾配降下法により学習を行った．

5.2 実験結果

5.2.1 劣化音声データを用いた話者認識モデルの構築

図 1 に，劣化音声を用い異なる音響特徴量 (MFCC, MFCC+F0) と手法 (GMM, i-vec) を用いて構築した話者認識モデルにおける，教師なし話者適応の結果 (メルケプストラム歪み，LF0 RMSE) を示す．適応データには

OFFICE と MEETING の劣化音声を用いたが，劣化の種類は既知としているため，同種の劣化音声を用いて構築された話者認識モデルを用い事後確率を計算し教師なし話者適応を行った．図 1 からわかるように，F0 に関する音響特徴量を話者認識モデルの構築に利用した場合，全ての手法でメルケプストラム歪み，LF0 RMSE の値が非常に高くなっている．これは，劣化音声から適切に基本周波数の抽出ができず，話者認識モデルの学習に悪影響を及ぼし，話者類似度の推定が適切に行われなかったためだと考えられる．また，音響特徴量として MFCC を用いた GMM-UBM と i-vector/PLDA を比較すると，i-vector/PLDA が全ての条件で良い評価となっている．

より詳細に結果を見るため，図 2 に，異なる SNR の適応データを用いた教師なし話者適応の結果を示す．図 2 から分かるように，GMM-UBM を用いた手法では劣化の強度が強いほどメルケプストラム歪み，LF0 RMSE の値は高くなっているが，i-vector/PLDA を用いた手法の場合，どの SNR の場合でもほぼ同等の結果となっている．OFFICE の劣化音声の場合でも同様の傾向が見られた．これらの結果から，i-vector/PLDA により，劣化音声から各学習話者に対する事後確率がロバストに計算されたことがわかる．

5.2.2 従来法との比較

図 3 に話者認識モデルの学習に劣化の無い高品質音声データを用いた場合と，劣化音声データを用いた場合の教師なし話者適応の結果を示す．適応データに劣化音声を用いる場合は，音響特徴量に MFCC，モデルに i-vector/PLDA を利用した結果のみを示している．図 3 から分かるように，理想的な条件と言える，話者認識モデルの学習と適応データの両方に劣化の無い高品質音声データ (CLEAN/CLEAN) を用いた手法と比較すると，適応データに劣化音声を用いた場合，どの手法でも若干メルケプストラム歪み，LF0 RMSE が高くなっていることが分かる．適応データに劣化音声を用いた手法間で結果を比較すると，劣化の無い音声データを用い話者認識モデルを構築した場合と比べ，劣化音声データを話者認識モデルの学習データとして用いることで，メルケプストラム歪み，LF0 RMSE が改善している．このことから，劣化音声から話者認識モデルを学習することで，劣化音声から事後確率を計算するのに適した話者認識モデルの構築が行われたことがわかる．

6. おわりに

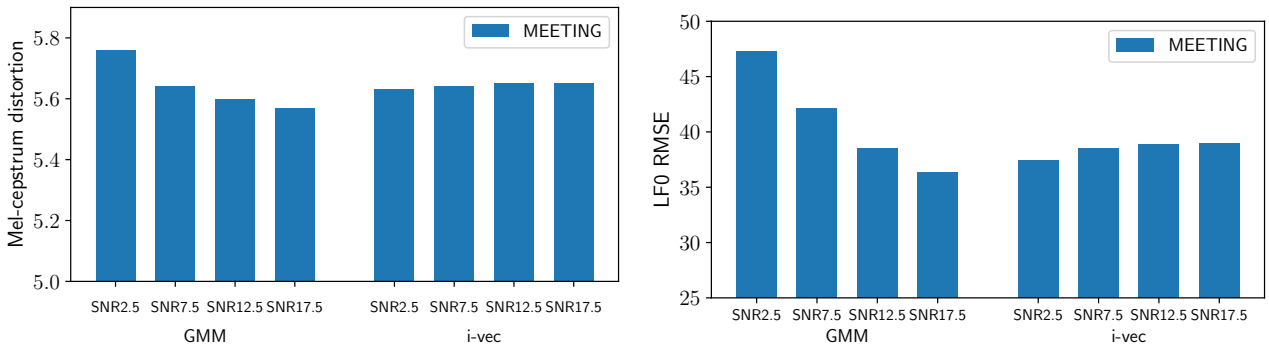
本論文では，DNN に基づく音声合成において，目標話者の音声が雑音や残響を含み劣化していることを想定し，このような劣化音声に対し，ロバストな教師なし話者適応について検討した．目標話者の劣化音声から適切に話者類似度が計算可能な話者認識モデルの構築を，雑音や残響を付与した劣化音声データを用いて行う．評価実験の結果から，劣化音声データを用いて話者認識モデルを構築することの有効性，GMM-UBM と比較し i-vector/PLDA を用いることで，DNN 音声合成システムの入力コードに適切な事後確率をロバストに計算可能であることがわかった．今後の課題としては，主観評価実験を実施すること，MP3 や AMR codec 等により圧縮された劣化音声の利用，より高品質な音声合成の実現のため，ボコーダを用いない波形生成における教師なし話者適応手法の検討が挙げられる．

7. 謝辞

本研究は株式会社オルツの助成を受けた．

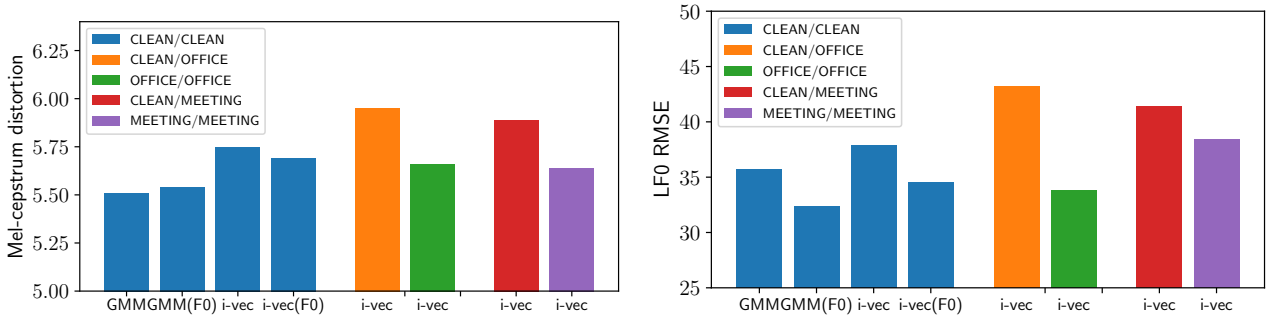
参考文献

[1] Hojo, N., Ijima, Y. and Mizuno, H.: An Investigation of DNN-Based Speech Synthesis Using Speaker Codes, *Proc. Interspeech* (2016).
[2] Arik, S. Ö., Diamos, G. F., Gibiansky, A., Miller, J., Peng, K., Ping, W., Raiman, J. and Zhou, Y.: Deep Voice 2: Multi-Speaker Neural Text-to-Speech, *CoRR*, Vol. abs/1705.08947 (online), available from ⟨http://arxiv.org/abs/1705.08947⟩ (2017).
[3] Taigman, Y., Wolf, L., Polyak, A. and Nachmani, E.: Voice Synthesis for in-the-Wild Speakers via a Phonological Loop, *CoRR*, Vol. abs/1707.06588 (online), available from ⟨http://arxiv.org/abs/1707.06588⟩ (2017).
[4] Wu, Z., Swietojanski, P., Veaux, C., Renals, S. and King, S.: A study of speaker adaptation for DNN-based speech synthesis, *Proc. Interspeech* (2015).
[5] Swietojanski, P. and Renals, S.: Learning Hidden Unit Contributions for Unsupervised Speaker Adaptation of Neural Network Acoustic Models, *Proc. SLT*, pp. 171–176 (2014).
[6] Fan, Y., Qian, Y., Soong, F. K. and He, L.: Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis, *Proc. ICASSP*, pp. 4475–4479 (2015).
[7] Luong, H.-T., Takaki, S., Henter, G. E. and Yamagishi, J.: ADAPTING AND CONTROLLING DNN-BASED SPEECH SYNTHESIS USING INPUT CODES, *Proceedings of ICASSP*, pp. 4905–4909 (2017).
[8] 高木信二，西村祥一，山岸順一：DNN 音声合成のための話者類似度に基づく教師なし話者適応，第 118 回音声言語情報処理研究会.
[9] Reynolds, D. A., Quatieri, T. F. and Dunn, R. B.: Speaker Verification Using Adapted Gaussian Mixture Models, *Digital Signal Processing*, Vol. 10, pp. 19–41 (2017).
[10] Kenny, P.: Bayesian Speaker Verification with Heavy-Tailed Priors, *Odyssey 2010* (2010).
[11] Thiemann, J., Ito, N. and Vincent, E.: The Diverse Environments Multi-Channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings, Vol. 133, p. 3591 (2013).
[12] Hadad, E., Heese, F., Vary, P. and Gannot, S.: Multi-channel audio database in various acoustic environments, *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 313–317 (online), DOI: 10.1109/IWAENC.2014.6954309 (2014).
[13] Han, K., Wang, Y., Wang, D., Woods, W. S., Merks, I. and Zhang, T.: Learning Spectral Mapping for Speech Dereverberation and Denoising, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 23, No. 6, pp. 982–992 (online), DOI:



(a) メルケプストラム歪み　(b) LF0 RMSE

図 2 異なる SNR の適応データを用いた教師なし話者適応の結果 (メルケプストラム歪み，LF0 RMSE)．SNR2.5，SNR7.5，SNR12.5，SNR17.5 の数字は SNR の値を示し，単位は dB である．また，ここでは話者認識モデルの学習，適応データに MEETING の劣化音声を用いた結果を示す．



(a) メルケプストラム歪み　(b) LF0 RMSE

図 3 話者認識モデルの学習に劣化の無い高品質音声データを用いた場合と，劣化音声データを用いた場合の教師なし話者適応の結果 (メルケプストラム歪み，LF0 RMSE)．GMM(F0) と i-vec(F0) は MFCC に加え F0 に関する音響特徴量も用い，話者認識モデルの構築を行った．また，ラベルは前半 (/の前) が話者認識モデルを学習した音声の種類，後半 (/の後) が適応データの音声の種類を表す．例えば，CLEAN/MEETING であれば話者認識モデルの学習に劣化の無い高品質音声データ，適応データとして MEETING の劣化音声が用いられたことを表す．

10.1109/TASLP.2015.2416653 (2015).
[14] Larchera, A., Lee, K. A. and Meignier, S.: An extensible speaker identification SIDEKIT in Python, *Proceedings of ICASSP* (2016).