Vol.2021-SLP-136 No.29
2021/3/4

短時間発話を用いた話者照合のための音声加工の効果に関する検討

宋 裕進[1,a)]　塩田 さやか[2,b)]　高道 慎之介[3,c)]　村上 大輔[4,d)]　松井 知子[4,e)]　猿渡 洋[3,f)]

概要：事前に登録された音声と入力された音声が同一話者のものであるかを判別するタスクである話者照合においては，発話データから deep neural network を用いて x-vector とよばれる話者表現を抽出したのち，probabilistic linear discriminant analysis (PLDA) を用いて識別を行う方法が近年の最先端技術として用いられてきた．しかし，用いられる発話データの長さが十分でない場合，x-vector に話者の情報が十分反映されず，照合の精度が安定しないという問題点がある．そこで本研究では，短時間音声からも話者情報をより頑健に抽出し照合精度を高めるために，登録発話や照合用発話に対して様々な音声加工を施すことを検討する．実験ではまず，複数の音声を連結することで長さを伸長させる方法や，複数の音声波形を重ね合わせる方法について識別率の挙動を確認した．実験結果から，発話の連結や重ね合わせによって識別率が向上したことを報告する．次に，照合用発話の量に制約がある場合に識別率を向上させる方法を検討するために，登録音声から一部を切り出して入力音声との連結・重ね合わせを行う場合や，waveform similarity overlap-add や phase vocoder などの信号処理的手法を用いて音声波形を伸長させる場合それぞれについて識別率の挙動を検証し，報告する．

Yujin Song[1,a)]　Sayaka Shiota[2,b)]　Shinnosuke Takamichi[3,c)]　Daisuke Murakami[4,d)]
Tomoko Matsui[4,e)]　Hiroshi Saruwatari[3,f)]

1. はじめに

話者照合とは，生体情報として音声を利用した生体証技術である．音声はマイクがあれば入手できることから，特殊なデバイスを用いず簡単に採取できる生体情報のひとつであり，スマートフォンやパーソナルコンピュータなどと親和性も高いことから，さまざまなアプリケーションの個人認証技術として話者照合を用いることができると期待される．そのため，実用に耐えうる話者照合技術の実現に向けた研究が近年活発に進められてきている．話者照合においては，事前に入力された登録発話が照合時に入力された照合用発話と同一人物によるものか否かを判定するため，音声データから話者固有の情報である話者表現を適切に分離することが重要である．近年では deep neural network (DNN) を用いた話者表現である x-vector に基づく手法 [1] が最先端技術として注目されている．

最先端である x-vector に基づく話者照合においても，登録発話や照合用発話の長さが非常に短い場合，抽出される話者の情報量が不十分となり，識別率が低下してしまうという問題が広く知られている [2]．この問題を改善するために，学習に用いるデータを増やすことで話者情報の頑健性を向上させるデータ拡張の手法が提案されてきている．例えば，DNN の学習用音声に複数種類のノイズを重畳する手法 [1]，x-vector 自体を深層生成モデルによって生成する手法 [3–5] などが挙げられる．しかしながら，これらの手法においても性能が不十分であったり，前提とするシステムが実用に適さないなど様々な問題が残っている．

本研究では，特に実用面を考慮し，登録発話や照合用発話の長さが限られているという制約の下で識別率を向上させる手法について検討する．そのために，発話長の短い登録発話や照合用発話を加工することによって話者情報の頑健性を増す手法に焦点をあてた．そこでまず，発話の長さ

1 東京大学工学部計数工学科
2 東京都立大学／統計数理研究所
3 東京大学 大学院情報理工学系システム情報学専攻
4 統計数理研究所
a) song-yujin139@g.ecc.u-tokyo.ac.jp
b) sayaka@tmu.ac.jp
c) shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
d) dmuraka@ism.ac.jp
e) tmatsui@ism.ac.jp
f) hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp

が具体的に照合性能にどの程度影響を与えるのかを確認するために，複数の照合用発話を連結した場合・重ね合わせた場合それぞれについて識別率の挙動を実証実験により確認した．どちらの場合でも識別率が向上するものの，一定以上の連結や重ね合わせを行うと識別率が打ち止めになることを報告する．次に，1 回の照合あたりの照合用音声の発話長が非常に短いという制約下で識別率を向上させる手法として，登録発話から一部を切り出して照合用発話との重ね合わせ・連結を行う方法や，waveform similarity overlap-add (WSOLA) [6] や phase vocoder [7,8] などの信号処理的手法によって発話の周波数を保ったまま発話長を伸長する方法について検討した．本実験では，これらの結果が識別性能にどのような影響を与えるのかを調査し，報告する．

2. X-vector に基づく話者照合

2.1 概要

X-vector に基づく話者照合手法とは，話者表現を抽出するためのモデルを DNN により構築し，ネットワークの埋め込み層の出力を特定話者の話者表現ベクトルであるとみなすものである．まず，表 1 に x-vector を抽出するための DNN の構成を示す．表 1 における T および N はそれぞれ DNN に入力する音響特徴量のフレーム数，データセットを構成する話者数を表す．DNN に入力する音響特徴量には発話から抽出した v_0 次元の音響特徴量を用いる．DNN の第 1 層から第 5 層は time-delay neural network (TDNN) [9] によって構成される．表 1 における TDNN1 から TDNN5 までの層は各フレーム t $(1 \leq t \leq T)$ ごとに並列に以下のように処理を行う．

TDNN1 区間 $[t-2,\ t+2]$ に含まれるフレームの MFCC を合わせた $5v_0$ 次元のベクトルが入力され，v_1 次元のベクトルが出力される．

TDNN2 フレーム $t-2,\ t,\ t+2$ における TDNN1 の出力を合わせた $3v_1$ 次元のベクトルが入力され，v_1 次元のベクトルが出力される．

TDNN3 フレーム $t-3,\ t,\ t+3$ における TDNN2 の出力を合わせた $3v_1$ 次元のベクトルが入力され，v_1 次元のベクトルが出力される．

TDNN4, TDNN5 フレームごとに独立に処理され，最終的に v_2 次元の出力ベクトルが出力される．

TDNN5 の出力は statistical pooling 層とよばれる第 6 層で統合される．この層では，各フレームの TDNN5 の出力である v_2 次元ベクトルの平均と標準偏差を連結した $2v_2$ 次元のベクトルが出力される．以降の層ではフレーム方向の情報は保持せず，最終的に N 種類の話者のいずれであるかを softmax 層で抽出する．全ての層でバッチ正規化を行い，活性化関数には rectified linear unit [10] を用いる．TDNN6 あるいは TDNN7 の出力である d 次元ベクトルは，話者を識別するために学習した DNN の中間的な出力であり，話者情報が埋め込まれたベクトルとして，x-vector と呼ばれている．

表 1 X-vector を抽出するための DNN の構成

層	入力フレーム	入力サイズ	出力サイズ
TDNN1	$[t-2,\ t+2]$	$5v_0$	v_1
TDNN2	$\{t-2,\ t,\ t+2\}$	$3v_1$	v_1
TDNN3	$\{t-3,\ t,\ t+3\}$	$3v_1$	v_1
TDNN4	$\{t\}$	v_1	v_1
TDNN5	$\{t\}$	v_1	v_2
stat-pooling		$T \times v_2$	$2v_2$
TDNN6		$2v_2$	d
TDNN7		d	d
softmax		d	N

2.2 Probabilistic linear discriminant analysis (PLDA)

X-vector に基づく話者照合では，照合時に照合用発話から抽出された x-vector と登録発話から抽出された x-vector との類似性を測り，その類似度をもとに同じ話者であるか否かを判別する．その際に用いられる確率モデルが probabilistic linear discriminant analysis (PLDA) である．PLDA では，話者 i $(1 \leq i \leq N)$ の発話 k $(1 \leq k \leq M_i)$ を表す話者表現 $\boldsymbol{x}_{ik}$ を，その生成過程を考慮せずに，話者潜在ベクトル $\boldsymbol{y}_i$ を平均，Φ_w を分散共分散行列とする正規分布から独立に生成されたベクトルであると仮定する．さらに，$\boldsymbol{y}_i$ の事前分布を各 i ごとに独立な平均 $\boldsymbol{m}$，分散共分散行列 Φ_b の正規分布で設定する．最尤法でモデルのパラメータ $\boldsymbol{m}$, Φ_w, Φ_b を学習したうえで，登録音声と入力音声が同じ潜在変数から生成される尤度 P_s と登録音声と入力音声が異なる潜在変数から生成される尤度 P_d の対数尤度比

$$L = \log \frac{P_s}{P_d} \tag{1}$$

を照合スコアとし，スコアが閾値よりも高い場合に，照合用発話と登録発話の話者が同じであると判断する．なお，登録発話が複数あった場合は，それらすべてを用いて対数尤度を計算することでより識別精度を高めることができる．

3. 認証データの頑健性向上のための音声加工

登録発話や照合用発話の発話長が非常に短い場合に，x-vector などの話者表現の抽出において話者の情報量が不十分となり照合性能が安定しないという問題があるが，実用上は登録発話や照合用発話の発話長が短いことが前提とみなす必要がある．そこで短い発話に加工を施すことにより，抽出された x-vector に含まれる話者情報を頑健なものにする手法を検討する．

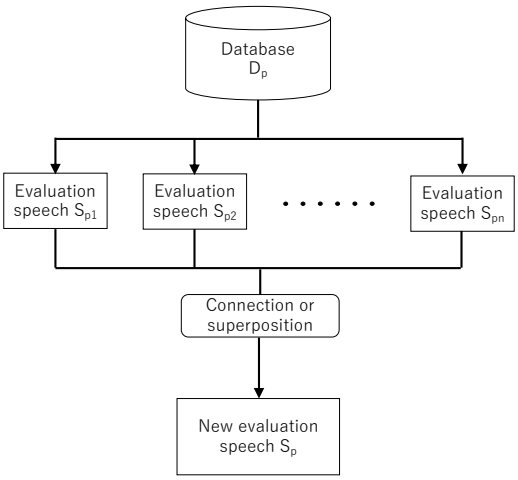

図 1 複数の照合用発話に対し連結・重ね合わせを行う手法.

3.1 発話の連結・重ね合わせ

複数の照合用発話を連結，あるいは重ね合わせする手法を検討する．手法の概要を図 1 に示した．話者を p とし，この手法では p の発話を蓄積したデータベース D_p を事前に用意しておく．D_p から n 個の発話 $S_{p1}, \cdots, S_{pn}$ を選び，これらに対して連結，重ね合わせ処理を施す．まず，$S_{p1}, \cdots, S_{pn}$ を連結する手法について述べる．2.1 節で述べた通り，DNN の statistical pooling 層では，前層の出力のフレーム方向の平均と標準偏差が出力されることでフレーム方向の情報が圧縮される．n 個の短時間発話音声 $S_{p1}, \cdots S_{pn}$ を連結した音声 S_p を作成することで，statistical pooling 層から出力される平均ベクトルの揺らぎが大きい問題を緩和する．具体的には，$S_{p1}, \cdots, S_{pn}$ それぞれから抽出される MFCC を連結した MFCC と S_p の MFCC は，音声連結部の近傍を除いておよそ等しいため，DNN に S_p を入力した場合に出力される平均ベクトルは，DNN に $S_{p1}, \cdots, S_{pn}$ を入力した場合に出力される n 個の平均ベクトルの平均におよそ等しい．したがって DNN に S_p を入力した場合には，$S_{p1}, \cdots, S_{pn}$ を入力した場合に比べ揺らぎの小さく安定した平均ベクトルを抽出可能である．これにより，S_p を話者 p の新たな照合用発話とすることで，最終的な話者情報を頑健に抽出できると期待される．

発話に含まれる話者情報を頑健にする別の方法として，複数の音声波形を重ね合わせる方法も考えられる．n 個の発話 $S_{p1}, \cdots, S_{pn}$ を重ね合わせたものを話者 p の新しい照合用発話 S_p とする．一般に多くの音声波形を重ね合わせた場合明瞭性は低下するうえ，音声は一定の確率分布に従う信号に収束するため，音声を多く重ね合わせると話者性はむしろ弱まってしまうと推測される．しかし，重ね合わせ数があまり多くない場合は明瞭性がそれほど下がらず，重ね合わせのない単一音声よりも各フレームの話者情報が頑健になる可能性がある．

3.2 登録発話の一部を切り出す手法

3.1 節で述べたような音声の連結・重ね合わせを用いた方法で識別率を高めることができても，照合用発話の量が限られている状況では，発話の連結や重ね合わせが十分にできないという問題がある．発話を登録する機会が複数あるような状況では，登録発話に含まれる話者の情報量は十分に多いと仮定することができる．そのため，登録発話から音声の一部区間を切除したとしても，照合性能に与える影響は小さいと考えられる．そこで，一度の話者照合ごとに登録発話から一部区間を切除し，その区間を照合用発話と連結する，あるいは重ね合わせることにより，照合用発話の情報量を増やし，照合性能を向上できる可能性がある．手順としてはまず，照合を行う前に登録発話から長さ d_1 の区間をランダムに選択して切除し，残りを新しい登録発話とする．次に，切除した長さ d_1 の区間から各照合ごとに長さ d_2 の区間をランダムに選択し，照合用発話の末尾に連結するか，照合用発話に重ね合わせする．もしも登録発話と照合用発話が同一人物によるものであれば，照合用発話に含まれる話者情報の頑健性は向上し，より同一人物であると識別しやすくなる．しかし，登録発話と照合用発話が同一話者のものでない場合は，異なる話者の発話を混合するため，照合用発話に含まれていた話者性が弱まると考えられる．このため照合性能は，登録発話と照合用発話が同一話者のものである場合の識別率の向上と，異なる話者のものである場合の識別率の低下のうちどちらが大きく寄与するかに依存すると考えられる．

3.3 信号処理的手法

登録発話や照合用発話の量が少ないという条件のもとで話者の情報量を増やすもう一つの手法として，発話音声を加工することで発話長を伸ばす手法を検討する．リサンプリングに基づく単純な音声波形伸長は，周波数スペクトルを低周波数方向に縮退させるため，伸長前の発話に含まれる話者情報が損なわれるという問題点がある．そこで発話音声の話者性を保つために，音声波形の周波数を変化させず長さのみを変更する信号処理的手法である WSOLA [6] および phase vocoder [7,8] を用いた．

3.3.1 WSOLA

WSOLA では，加工前の音声波形を短い部分波形に分割し，窓掛け処理を施したのち新たな間隔で再配置することで音声長を伸長する．まず，加工前の音声波形 $x(t)$ を幅 N，シフト幅 H_a の矩形窓で切り出すことで部分波形 $x_m(t)$ $(m \in \mathbb{Z})$ を構成する．次に，各 $x_m(t)$ にフェードイン・フェードアウトの窓掛け処理を施した波形 $x'_m(t)$ を時間軸上に間隔 $H_s + \Delta_m$ で再配置する．すなわち，

$$y(t) = \sum_{m \in \mathbb{Z}} x'_m(t - mH_s - \Delta_m) \qquad (2)$$

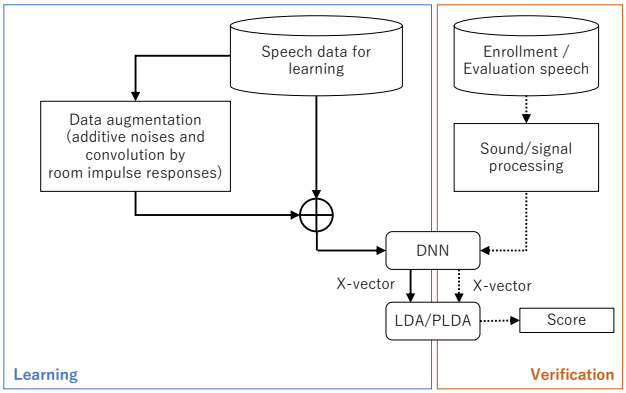

図 2 x-vector に基づく話者照合システムのアーキテクチャ．

とする．ここで繋ぎ目における不連続性を小さくするため，Δ_m は隣り合う部分波形同士の相関関数が最も高くなるように定める．$H_s > H_a$ の場合，得られた波形 $y(t)$ は元の波形 $x(t)$ よりも長くなる．

3.3.2 Phase vocoder

Phase vocoder に基づく波形伸長でも WSOLA と同様に，加工前の音声波形 $x(t)$ を幅 N，シフト幅 H_a の矩形窓で切り出し部分波形 $x_m(t)$ $(m \in \mathbb{Z})$ を構成する．そして各 $x_m(t)$ のフーリエ変換により得られるスペクトル $X_m(\omega)$ を，隣り合う部分波形の位相に不連続性が生じないよう修正してから逆フーリエ変換して修正波形 $x'_m(t)$ を得る．この修正波形を時間軸上に間隔 H_s で再配置する．すなわち，

$$y(t) = \sum_{m \in \mathbb{Z}} x'_m(t - mH_s) \qquad (3)$$

とする．$H_s > H_a$ の場合，得られた波形 $y(t)$ は元の波形 $x(t)$ よりも長くなる．

4. 実験

4.1 実験条件

本実験で用いた話者照合システムのアーキテクチャを図 2 に示した．DNN の学習に用いたデータは，音声コーパス VoxCeleb1,2 [11,12] に含まれる 7,245 話者による発話数 1,245,525 の発話データである．発話データのサンプリング周波数はすべて 16kHz であった．さらに，noise, music, babble の 3 種類の背景音からなるノイズのコーパス MUSAN noise [12] と，室内インパルス応答を含むコーパス room impulse responses noise [13] を用いて学習用発話にノイズを加えたものも図 2 のように学習データに加えた．このデータ拡張の結果，学習に用いる発話数はもとの 5 倍になっている．

図 2 における DNN の構成は表 1 において $v_0 = 30, v_1 = 512, v_2 = 1500, d = 512$ としたものを用いた．音響特徴量には $v_0 = 30$ 次元の mel frequency cepstral coefficient (MFCC) を用いた．DNN から抽出された $d = 512$ 次元の x-vector を，事前に linear discriminant analysis (LDA) で 200 次元に縮約し，この 200 次元ベクトルを PLDA による照合に用いた．なお，PLDA を学習するために用いる発話データには VoxCeleb1,2 の 1,245,525 発話すべてを用いたが，ノイズや室内インパルス応答によるデータ拡張は施していない．

登録発話と照合用発話は，音声コーパス RedDots Dataset [14] のサブセットである Part 4 に含まれる男性 49 話者，女性 13 話者から構成される 62 話者の発話データ 15,343 発話を用いた．コーパスには，実際の話者照合に用いることを想定して数字や慣用句などの短く単純なフレーズが収録されている．発話データのサンプリング周波数はすべて 16kHz であり，1 発話あたりの平均発話長は 3.359 秒であった．このうち男性 35 話者，女性 6 話者から構成される 41 話者の発話データは登録データと入力データの両方に用い，残り 21 話者のデータは入力データにのみ用いた．各話者ごとに複数の異なる登録発話データ 2 ～ 10 発話が存在する．また，PLDA を用いた照合の識別率の評価には，本人受入率と他人棄却率が等しくなる場合の誤認識率である equal error rate (EER) を採用した．本実験のベースラインとして，各試行ごとの照合用発話は単一の発話のみで，もとの音声に加工を施さないという条件で照合を行った．ベースラインの登録発話については，話者ごとに全ての登録発話を用いて照合スコアを計算する場合をベースライン 1，登録発話を話者ごとに 1 つだけ選択して計算する場合をベースライン 2 として実験を行った．

表 2 照合用発話の連結による識別率の挙動

連結数	平均発話長 [s]	EER[%]
1	3.359	3.258
2	6.795	0.9949
3	10.25	0.4533
4	13.64	0.3153
5	17.11	0.2063
6	20.54	0.1397
7	24.00	0.09803
8	27.45	0.07901
9	30.82	0.08413
10	34.28	0.05852

4.2 照合用発話の連結・重ね合わせ

ベースライン 1 の EER は 3.258%であり，ベースライン 2 の EER は 8.457%であった．このことから登録発話について，短時間の発話を 1 つしか用いなかった場合話者情報が不安定になり，大きく照合性能が低下することがわかる．次に，ベースライン 1 と同様に話者ごとに全ての登録発話を用いて照合スコアを計算するという条件のもとで，複数の照合用発話を連結して発話長を伸ばしていった場合に識別率はどのような挙動を示すのかを検証した結果を表

表 3 照合用発話の重ね合わせによる識別率の挙動

重ね合わせ数 n	EER[%]
1	3.258
2	2.102
3	3.056
4	3.567

表 4 登録発話の一部を切り出して照合用発話に連結・重ね合わせした場合

d（連結）	EER[%]	r（重ね合わせ）	EER
l	18.06	1	20.66
$l/2$	12.70	1/2	13.30
$l/4$	8.829	1/4	9.931

2 に示す．平均 3 秒程度の照合用発話について，連結数が増えた場合 3 発話程度までは急激に識別率が上昇するが，それ以上連結数が増加するにつれて識別率が打ち止めになることがわかる．

次に，ベースライン 1 と同様に話者ごとに全ての登録発話を用いて照合スコアを計算するという条件のもとで，複数の照合用発話を重ね合わせた場合の認識精度の挙動を検証した．結果は表 3 に示した．表 3 における n は重ね合わせた発話の個数を表す．$n = 2$ のときベースライン 1 と比較して識別率が上昇したが，これ以上重ね合わせ数を増やした場合 $n = 2$ に比べて識別率は低下し，$n = 4$ 以降ではベースラインよりも識別率が下回った．実験により，2 発話の重ね合わせにより各フレームに含まれる話者情報は増加し，連結した場合ほどではないものの識別率は向上することが明らかになったが，3.1 節で述べたように，多くの発話を重ね合わせれば重ね合わせるほど音声の明瞭性や発話に含まれる話者性は弱まり，識別率が低下してしまうことも確認できた．

4.3 実験結果：登録発話の一部を切り出す手法

ベースライン 1 と同様に登録発話と照合用発話を用意した後，3.2 節で述べた手法を用いて登録発話から一部を切り出して用いる手法を検証した．登録発話から切除する区間の長さ d_1 は照合用発話の長さの 1/4 とした．照合用発話の長さを l とし，$d_2 = l/4,\ l/2,\ l$ の 3 通りに定めて実験をおこなった．同様に，今度は $d_2 = l$ として，切り出した区間の音量を r 倍に変更して入力音声と重ね合わせた．$r = 1/4,\ 1/2,\ 1$ として実験を行った 結果を表 4 に示した．いずれの場合においても，わずかでも入力音声に登録音声が混ざると，3.2 節に述べた通り登録発話と照合用発話が違う話者のものであった場合に登録発話の話者性が損なわれ，照合性能が大きく低下してしまったと考えられる．

4.4 実験結果：信号処理的手法

最後に，ベースライン 1 と同様に話者ごとに全ての登録発話を用いて照合スコアを計算するという条件のもとで，WSOLA および Phase Vocoder を用いて入力音声と登録音声の双方の発話長を r 倍に伸長して照合を行った．$r = 1.125,\ 1.25,\ 1.5$ の 3 通りに定めて実験を行った．結果は表 5 に示した．音声に含まれる話者の情報量はあまり変わらないが，音声信号を処理する段階で音声信号が一定の割合で破壊されたことでむしろ識別率が低下してしまったと推測される．

表 5 信号処理的手法を用いて発話を伸長した場合

r[Phase Vocoder]	EER[%]	r[WSOLA]	EER[%]
1.125	7.721	1.125	3.451
1.25	7.837	1.25	3.652
1.5	8.155	1.5	4.184

5. おわりに

本稿では，音声の連結や重ね合わせによって識別率が向上し，連結した場合については次第に識別率が飽和することを確認できたが，照合用発話が限られているという制約の下ではベースラインよりも識別率が低下した．

今後の課題としては，照合用発話が限られている条件で識別率を向上させる手法を検討するとともに，より人為的なパラメータの少ない手法のために，識別率と発話長との関係が定量的にどのように評価できるのかも検討する必要があると考えられる．

謝辞

本研究の一部は，ROIS-DS-JOINT (023RP2020)，セコム財団挑戦的研究助成の助成を受けたものである。

参考文献

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust DNN embeddings for speaker recognition,” in *Proc. ICASSP*, 2018.

[2] H. Zeinali, K. A. Lee, J. Alam, and L. Burget, “Short-duration Speaker Verification (SdSV) challenge 2020: the challenge evaluation plan,” *arXiv preprint arXiv:1912.06311*, 2020.

[3] Y. Yang, S. Wang, M. Sun, Y. Qian, and K. Yu, “Generative adversarial networks based x-vector augmentation for robust probabilistic linear discriminant analysis in speaker verification,” in *Proc. ISCSLP*, 2018.

[4] Z. Wu, S. Wang, Y. Qian, and K. Yu, “Data augmentation using variational autoencoder for embedding based-speaker verification,” in *Proc. INTERSPEECH*, 2019, pp. 1163–1167.

[5] S. Shiota, S. Takamichi, and T. Matsui, “Data augmentation with moment-matching networksfor i-vector based speaker verification,” in *Proc. APSIPA*, 2018, pp. 345–349.

[6] W. Verhelst and M. Roelands, “An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech,” in *Proc. ICASSP*, 1993.

[7] J. L. Flanagan and R. M. Golden, “Phase vocoder,” *Bell*

Syst. Tech. J., no. 45, pp. 1493–1509, 1966.

[8] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Trans. Speech Audio Process.*, no. 7, pp. 323–332, 1999.

[9] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural networks," in *Proc. IEEE*, 1989, pp. 328–339.

[10] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105.

[11] A. Nagrani, J. S. Chung, , and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[12] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2:deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[13] P. Wang, Y. Qian, F. K. Soong, L. He, and H. Zhao, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. ICASSP*, 2017, pp. 5220–5224.

[14] K. Lee, A. Larcher, G. Wang, P. Kenny, N.Brümmer, D. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma, H. Li, T. Stafylakis, J. Alam, A. Swart, and J. Perez, "The reddots data collection for speaker recognition," in *Proc. INTERSPEECH*, 2015, pp. 2996–3000.