Vol.2021-NL-251 No.6
Vol.2021-SLP-139 No.6
2021/12/1

クラウドソーシングを用いた結果の検証による話者照合性能の改善

井手 悠太[1]　斎藤 奨[1,2]　中野 鐵兵[1,2]　小川 哲司[1]

概要：話者照合システムによる照合結果をクラウドソーシングを用いて人手で再検証することで，照合誤りを削減することを試みた．照合結果のうち信頼性が低いものを人手で再検証することで，効率的に誤りを削減できる可能性がある．その際に，適切に誤りが訂正され，正解であったものが誤訂正されないことが理想的であるが，クラウドワーカの話者照合能力や悪意のあるワーカの影響で，それが実現できるかは明らかになっていない．そこで，実際に自動話者照合システムによる照合結果の再検証をワーカに依頼し，誤受理率と誤棄却率の改善可能性について調査を行った．Amazon Mechanical Turk 上で，資格試験により一定以上の話者照合能力が保証された 426 名のワーカに対して 256 件の発話音声対の照合を依頼したところ，話者照合システムによる照合結果に対して誤棄却を増加させずに誤受理を大幅に削減できることが明らかになった．

キーワード：クラウドソーシング，MACE，話者照合

Improving speaker identification performance by crowd-assisted verification of results

Yuta Ide[1]　Susumu Saito[1,2]　Teppei Nakano[1,2]　Testuji Ogawa[1]

1. はじめに

x-vector をはじめとする深層話者埋め込みは，話者照合システムの性能を大幅に向上させ，現在の標準的な特徴抽出技術となっている [1], [2], [3], [4], [5]．この方式を基礎として，話者埋め込み [6], [7], [8], [9], [10] やスコアリング [5], [11], [12] について多くの検討がなされており，その有効性が明らかになっている．しかし，多様な雑音が混在するような実環境においては，照合性能は依然として改善の余地がある [13], [14]．話者照合が機械学習に基づいている以上，あらゆる環境において頑健に完璧な照合を達成することは現実的ではない．それに対し，話者照合システムの誤りを人手で訂正することができれば，特にシステムにとって未知の環境において，運用可能なレベルの性能を達成できる可能性がある．

人手による話者照合については，クラウドソーシングの利用可能性について検討がなされている [15]．ここでは，悪意のあるワーカや話者照合能力の低いワーカの回答を除くことで，音声からの話者照合が十分可能であること，および効率的に話者アノテーションを行うためのユーザインタフェース（UI）設計が明らかになっている．他にも，合成音声の品質評価 [16], [17] や音声からの感情推定 [18], [19] に対するクラウドソーシングの利用可能性が示唆されている．また，音声アノテーションについても，ワーカの集中力を低下させないよう複雑なタスクを分解するなどの条件設定が必要であるものの，十分な効果があることがわかっている [20]．

そこで，本研究では，話者照合システムによる照合結果のうち信頼性の低い入力発話対について，クラウドソーシングを用いて再検証を行うことで照合誤りを削減するこ

[1] 早稲田大学 基幹理工学部 情報通信学科
Department of Communications and Computer Engineering, Waseda University
[2] 知能フレームワーク研究所
Intelligent Framework Lab.

とを試みる．その際，照合誤りが正しく訂正され，正解が誤って修正されないことが理想的であるが，クラウドワーカの話者照合能力や悪意のあるワーカ（スパマー）の存在により，それが実現できるかは明らかではない．また，話者照合システムは，認証における信頼性担保という観点では誤受理率（false acceptance rate; FAR）を最小化することが望ましく，ユーザビリティの担保という観点では誤棄却率（false rejection rate; FRR）の増加を抑えることが望ましい．以上を考慮して，クラウドワーカによる照合結果の検証性能について，FRR と FAR に焦点を当てて調査を行う．ここでは，FRR の増加を抑えながら FAR を低減できることを期待する．

話者埋め込みとして x-vector，スコアリングに確率的線形判別分析（probabilistic linear discriminant analysis; PLDA）を用いた話者照合システムを対象とし，照合結果のうちスコアの絶対値が低い入力発話対の再検証をクラウドワーカに依頼する．この際，事前に実施した話者照合能力を判定する資格テスト [15] に合格したワーカにのみ検証を依頼することで，悪意のあるワーカや音声聴取能力の低いワーカの回答を取り除く．また，収集した回答の信頼性を向上させるために，1) 回答の多数決，2) 回答の偏りを考慮した多数決，3) ワーカの回答戦略モデルを用いて真のクラスを推定する multi-annotator competence estimation（MACE）[21], [22] のいずれかを適用し，最終的な検証結果を得る．本実験では，クラウドソーシングを大規模かつ継続的に実施するための開発環境である Tutti 上でマイクロタスク設計および運用を行う．

本稿の主たる貢献は以下の 2 点である．

(1) 話者照合における標準的なデータセットである VoxCeleb データにおいて，x-vector を用いた話者照合システムによる照合結果をクラウドワーカが検証することによる性能改善，および改善の内容に関する知見を与える．
(2) 回答の多数決と MACE を用いた場合を比較することで，クラウドソーシングで収集した回答の信頼性を高めるための有効な方法に関する知見を与える．

本研究で得られた知見は，頑健な話者照合システムの構築において有用であり，自動話者照合システムと人手による話者照合の連携のための基礎となることが期待できる．

本稿の構成は以下の通りである．2 では，本研究で利用する要素技術について述べる．3 では，クラウドワーカによる照合結果の検証実験およびその結果について述べる．最後に，4 でまとめを述べる．

2. 要素技術

本研究の要素技術として利用するクラウドソーシングを利活用するための開発環境と，クラウドソーシングにより得た回答の信頼性を向上させるためのクラウドワーカの回答戦略のモデル化について述べる．

2.1 大規模クラウドソーシングを継続的に利用するための環境

本研究では，クラウドソーシング用マイクロタスク UI の設計に Tutti[*1] を利用した．Tutti とは，アノテーション作業をマイクロタスクとして外注するためのウェブ UI の設計を容易に行える環境である．クラウドソーシングを用いて大規模にアノテーション作業を行う際は，多くの場合，同一 UI 上で異なるデータを出し分けたり，多数のワーカの回答を収集したりする仕組み等，多くのシステム実装が求められる．そのため，実験完了までに膨大な時間を必要とするという問題点がある．一方 Tutti では，実際にラベルを収集したいデータのこと以外を気にかける必要が無くなるため，実験に要する時間を大幅に削減できる．

本実験で Tutti を用いる主な利点としては以下が挙げられる．

- 異なるデータを読み込む機能を持つウェブページの雛形が提供されており，若干の UI 設計変更と読み込むデータのアップロードだけでデータ収集の準備をおおよそ完了できる．
- 複数種類のウェブページの遷移図を設計する機能により，同一マイクロタスク内で同一 UI でのラベリング作業を反復したり，条件によって別の UI を出し分けたりするような複雑なタスクを設計できる．
- ワーカへのタスク自動割当機能により，提示するデータごとの目標収集回答数に合わせた適切なマイクロタスク外注を行うことができる．
- GUI コンソール上や API を用いて，収集したワーカの回答を即時に確認できる．

2.2 クラウドワーカの能力・適正の推定

クラウドソーシングによるアノテーションの信頼性を担保するために，複数のワーカから得た回答の多数決により最終的なクラスを決定することが一般的である．本研究では，そのような回答の多数決に加え，複数のワーカの回答を用いてワーカの能力・適正を推定しながら真のクラスを推定する MACE も用いた．

MACE では，クラウドワーカの回答について以下のような生成過程を考える．タスク i に対する真値（クラスのインデックスなど）T_i は一様分布から生成される．クラウドワーカ j がタスク i において誠実に回答しているかどうかを表す 2 値変数 S_{ij} は，クラウドワーカごとにパラメータ θ_j を持つベルヌーイ分布から算出される．ここで，クラウドワーカが誠実に回答している場合（$S_{ij} = 0$），ワーカの回答 A_{ij} は真値 T_i と一致すると仮定する．また，

[*1] https://www.tutti.ai/

クラウドワーカが誠実に回答していない場合 ($S_{ij} = 1$)，ワーカの回答 A_{ij} は真値 T_i に依存せず，パラメータ ξ_j を持つ多項分布から生成されるとする．このとき，ワーカの回答 A_{ij} は観測可能であるが，T_i および S_{ij} は観測不可能である．また，θ_j はクラウドワーカ j の信頼性を表しており，ξ_j はクラウドワーカ j が不誠実な回答をする際の振る舞いを表している．

依頼するタスクの数を N, 回答したクラウドワーカの数を M とすると，θ_j および ξ_j は，変分 EM アルゴリズムにより，以下の尤度を最大化することで求めることができる．

$$P(A;\theta;\xi) = \sum_{T,S}\left[\prod_{i=1}^{N} P(T_i)\prod_{j=1}^{M} P(S_{ij};\theta_j)P(A_{ij}|S_{ij},T_i;\xi_j)\right]$$

クラウドワーカの能力を示す θ_j , 不誠実な回答をする際の振る舞いを表す ξ_j が推定されたのち，それらを用いてタスクごとの回答ラベルが L 個あり，各々のタスク i の回答から回答ラベル l ごとの事後確率 P_{il} を以下の式で求める．

$$P_{il}(A;\theta;\xi) = \frac{1}{L}\prod_{j=1}^{M}(\theta_j \cdot \xi_j(l) + G_{il}\cdot(1-\theta_j))$$

この時，G_{il} は $A_{ij} = l$ の時に 1 となり，それ以外で 0 となる関数とする．この P_{il} を用いてエントロピー H_{il} を求める．

$$H_{il} = -\frac{P_{il}}{\sum_{l=1}^{L} P_{il}}\log\frac{P_{il}}{\sum_{l=1}^{L} P_{il}}$$

このエントロピー H_{il} が最大となる回答ラベルがタスクの真値であると推定される．

MACE は，言語アノテーションのために設計されたが，音響アノテーションにおいても高い効果を発揮することが明らかになっている [23].

3. 照合結果の検証実験

照合誤りが生じている可能性が高い発話音声対をクラウドソーシングによって人手で照合することで，照合誤りが削減されるか調査を行った．

3.1 音声データ

VoxCeleb1 および VoxCeleb2 を用いて実験を行った [24]．VoxCeleb1 および VoxCeleb2 は，各々学習データ（train）と評価データ（test）から成る．本実験では，VoxCeleb1 の train と VoxCeleb2 の train および test を用いて話者照合システムを構築し，VoxCeleb1 の test を用いて評価ならびにクラウドソーシングによる再検証を行った．

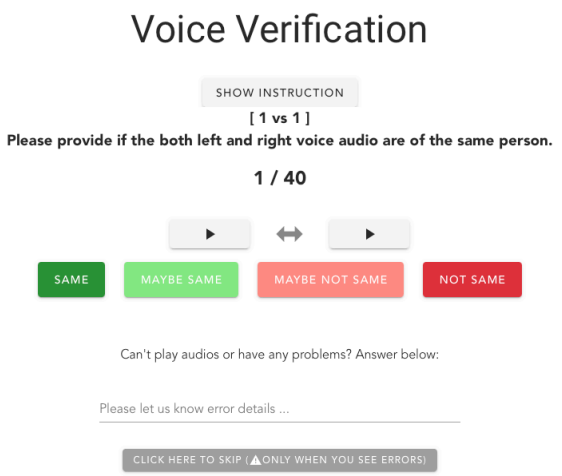

図 1 クラウドワーカの作業画面．2 つの発話音声を視聴し，それが同じ話者であるか，異なる話者であるかをその確信度とともに回答する．40 の発話対の照合を最小のタスクセット（HIT）とした．

3.2 話者照合システム

実験で使用した話者照合システムは，kaldi のレシピ[*2] を用いて作成した．つまり，Time delay neural network（TDNN）を用いて話者情報を埋め込んだ x-vector を特徴ベクトルとして用い，PLDA [25] により入力発話対の類似度を算出した [2]．このとき，前述の学習データを musan データ [26] を用いて拡張した．

3.3 クラウドソーシングによる照合結果の検証

クラウドワーカに依頼する照合タスクの概要およびその設計について述べ，クラウドワーカから取得した回答を集約し，照合の信頼性を向上させるための処理について説明する．

3.3.1 照合タスクの概要

図 1 に，クラウドワーカの作業画面を示す．クラウドワーカには発話音声のペアが提示され，各々に対応する再生ボタンをクリックすることで数秒の音声を視聴できる．視聴後，同一話者（same），おそらく同一話者（Maybe same），おそらく異なる話者（Maybe not same），異なる話者（Not same）のうちいずれかを選択させることで，検証結果を得た．このとき，各音声を最低 1 回再生するまでは回答を終了できないようにした．

クラウドワーカに依頼するタスクセットの最小単位を HIT と呼ぶ．本実験では，照合が容易な 8 つの発話対と，比較的難しい 32 の発話対から成る計 40 対の照合を 1 HIT とした．32 の発話対は 8 種類用意しているため，HIT は 8 種類となる．本実験では，1 種類の HIT に対して 15 人のワーカの回答を収集した．なお，事前に話者照合能力を測る資格試験を実施し [15]，全問正解したワーカのみを対

[*2] https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2

象として回答を収集した．

3.3.2 照合タスクの設計

クラウドワーカによる検証用データセットの構築方法，すなわち，ワーカが照合する発話対の選定方法について述べる．話者照合システムによる照合結果のうち PLDA スコアの絶対値が小さいものは，対応する入力発話対による照合の信頼性が低いことを意味する．このような発話対に対しては，照合誤りが生じている可能性が高い．実際，評価データに対する PLDA スコアは −130 から 75 の間に分布しているが，−9 から 9 の間に照合誤りのうち 85% が含まれていた．そこで本実験では，PLDA スコアの絶対値が 9 未満の発話対を照合の信頼性が低い入力とみなし，クラウドワーカによる検証の対象とした．

発話対は，PLDA スコアの絶対値がばらつくように選択した．ただし，スコアの絶対値が 0，3，5，8 の 4 種類を基準として，各基準値に対し，基準値以上最小の値から順番に対応する発話対を選択した．また，発話対の属性として，

- 発話対の性別が男性同士か女性同士か，
- 発話対が同じ話者によるものであるか否か，
- 話者照合システムが発話対を同じ話者と予測したか否か，

の組み合わせとして $2^3 = 8$ 種類を想定した．この発話対の属性全てを同数ずつ含むデータセットの話者照合システムによる照合結果における FRR および FAR は 0.5 となる．これら，スコア絶対値の基準値 4 種類，発話対の属性 8 種類の組み合わせとして，$4 \times 8 = 32$ 種類の発話対を 8 パターン得た．したがって，照合の信頼性が低い発話対としては，計 $32 \times 8 = 256$ 種類が検証対象となる．一方，照合が容易な発話対は，PLDA スコアの絶対値が大きく，かつ正解している発話対から選択した．

3.3.3 複数のワーカによる回答の集約

本実験では，[15] による知見に基づき，各発話対に対して 7 人のクラウドワーカによる回答の多数決を行う．その際，回答を収集するワーカの数が少ないと，特定のワーカの照合能力によって最終的な検証結果が左右されてしまう可能性がある．そのため，多数決を行う人数よりも多いワーカから回答を収集したうえで，7 人分の回答を選択する全ての組み合わせについて多数決を行うことで，特定のワーカの照合能力の影響を低減する．実際には，各発話対に対して 15 人分のクラウドワーカの回答を収集し，そのうち 7 人分の回答を選択する 6435 通りについて検証性能を調査した．

本研究では，この 6435 通りについて，以下の処理を行うことで検証結果の信頼性を向上させることを試みた．

(1) 複数ワーカによる回答の単純多数決（50% 以上のワーカが合意する場合）
(2) 複数ワーカによる回答の合意数を考慮した多数決（例えば，70% 以上のワーカが合意する場合）

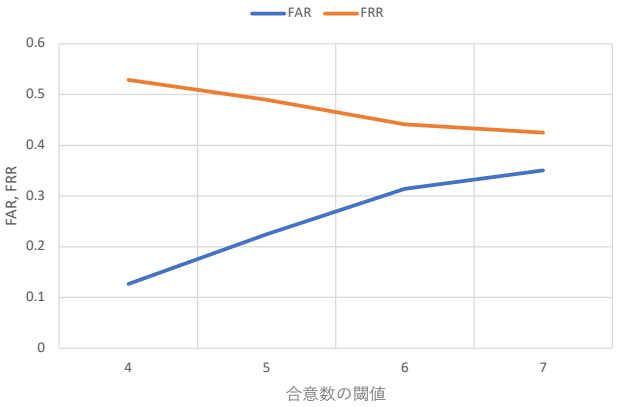

図 2 最低合意数（同一回答したワーカ数の下限）での多数決により検証を行ったときの誤受理率（FAR）と誤棄却率（FRR）．合意数を 5 以上とした多数決により FRR の増加を抑えつつ，FAR を最小化できる．

(3) 複数ワーカによる回答を用いた MACE [21]

ある発話対に対して同一の回答を行ったワーカの数（合意数）が多ければ，多数決の結果は信頼性が高い．そこで，合意数について閾値を設定し，合意数が閾値未満である発話対については話者照合システムの回答をそのまま採用し，閾値以上の合意数が得られた発話対についてはワーカの回答を採用する．この閾値を全ワーカのうち 50% とすれば単純多数決と等価であり，閾値を高くすれば信頼度が高い場合のみワーカの回答を採用することを意味する．

3.4 検証実験結果

本実験では，8 種類の検証データセットの各々を，話者照合システムの FAR と FRR がともに 0.5 となるように構築した．そこで，クラウドソーシングで収集した回答を用いた修正により，FAR と FRR を 0.5 よりどの程度低減できるかに着目して調査を行った．このとき，FRR の増加を抑えながら，FAR を削減できることが望ましい．

3.4.1 合意数と FAR および FRR の関係

図 2 に，複数のワーカの合意数を考慮した多数決により検証を行ったときの FAR と FRR を示した．図は，最低合意数（同一の回答を与えたワーカ数の下限）と FAR および FRR の関係を表している．このとき，最低合意数 N とは，N 人以上のワーカが同一の回答を与えたことを意味する．図より，最低合意数に依らず FAR は話者照合システムが与える 0.5 を下回った．これより，複数のワーカによる回答の多数決により FAR の削減が可能であることがわかる．一方で，FRR は最低合意数が 5 以上の場合に話者照合システムが与える 0.5 を下回った．このとき，FAR と FRR は合意数についてトレードオフの関係にある．つまり，必要な合意数が小さいほど FAR は低く，FRR は高くなり，その逆に必要な合意数が大きいほど FAR は高く，FRR は低くなった．そこで，FRR が 0.5 を下回る最小の

表 1 単純多数決，合意数を考慮した多数決 （70%以上のワーカの合意が必要），MACE を用いた場合の，正解維持，訂正成功，訂正失敗，誤訂正の発話対の割合の平均値．

	話者照合システム	ワーカによる検証	単純多数決	合意数を用いた多数決（閾値 5）	MACE
正解維持	正解	正解	0.45	0.54	0.46
訂正成功	不正解	正解	0.22	0.10	0.22
訂正失敗	不正解	不正解	0.18	0.30	0.18
誤訂正	正解	不正解	0.15	0.06	0.14

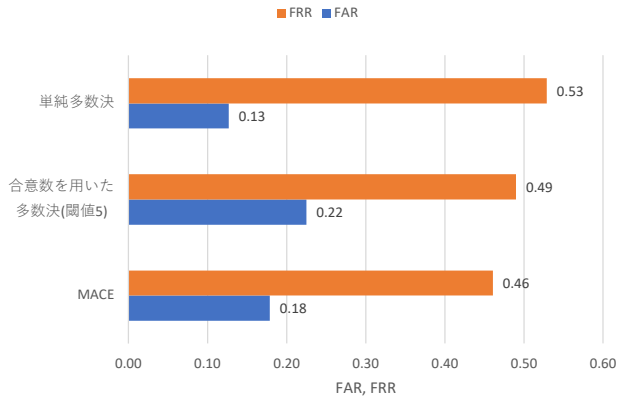

図 3 単純多数決，合意数を考慮した多数決 （70%以上のワーカの合意が得られた場合），MACE を用いてワーカの回答を集約したときの誤受理率 （FAR） および誤棄却率 （FRR）．

合意数である 5 を閾値として採用した場合 （70% 以上のワーカの合意が得られた場合） の検証結果を，合意数を考慮した多数決として，単純多数決，MACE と比較した．

3.4.2 複数のワーカによる回答の集約の効果

図 3 に，クラウドワーカから収集した回答を用いて単純多数決，合意数を考慮した多数決，MACE を行ったときの FAR および FRR を示した．単純多数決を用いた場合，FAR は最も削減されたものの FRR は劣化した．一方で，合意数を考慮した多数決と MACE を用いた場合，FAR と FRR をともに削減でき，MACE の方が誤りの削減度合いが大きいことがわかる．具体的には，MACE により，FRR を 7.9%，FAR を 64.3% 削減できることがわかった．認証の信頼性担保という観点では FRR の削減よりも FAR の削減が重要であることから，この結果は，クラウドワーカによる検証が理想的に機能したことを示唆している．

3.4.3 ワーカによる検証が照合結果に与える影響

クラウドワーカによる検証では，話者照合システムによる照合結果に誤りがあれば正しく訂正し，結果が正しければ誤って修正しないことが望ましい．そのような，クラウドワーカによる検証が照合結果に与える影響について調査を行った．このときの調査対象は以下の 4 通りである．

- **正解維持**：話者照合システムで正解した発話対に対し検証においても正しく照合できた．
- **訂正成功**：話者照合システムで誤った発話対に対し検証により正しく訂正できた．
- **訂正失敗**：話者照合システムで誤った発話対に対し検証により正しく訂正できなかった．
- **誤訂正**：話者照合システムで正解した発話対に対し検証により誤って修正した．

表 1 に，単純多数決，合意数を考慮した多数決，MACE を用いた場合の正解維持，訂正成功，訂正失敗，誤訂正の発話対の割合の 6435 通りの平均値を示した．70%以上のワーカの合意が必要な多数決 （合意数を考慮した多数決） では訂正の機会が減少するため，正解維持や訂正失敗の割合が必然的に高くなる．また，検証により悪影響を及ぼす割合 （訂正失敗と誤訂正の和） は，$0.30 + 0.06 = 0.36$ である．同様に，検証により悪影響を及ぼす割合は，単純多数決を用いた場合は $0.18 + 0.15 = 0.33$，MACE を用いた場合は $0.18 + 0.14 = 0.32$ であり，MACE を使用した場合に，検証における悪影響が最も少ないことがわかる．

4. まとめ

本研究では話者照合システムの照合結果をクラウドソーシングを用いて再検証することで，誤りを削減することを試みた．資格試験を通過したクラウドワーカにのみ回答を依頼し，クラウドワーカの回答の信頼性を向上させて話者照合システムの照合結果を検証することで FRR の増加を抑えながら FAR を大幅に削減することができることが明らかになった．また，単純多数決及び合意数を用いた多数決を使用した検証と比較して MACE を使用した検証が理想的に誤りを訂正できる．

本研究は，頑健な話者照合システムの構築にために自動話者照合システムとクラウドソーシングの連携のための基礎となることを期待している．しかし，話者照合システムにとって不利な環境での，照合性能の改善可能性については明らかになっていない．話者照合システムの性能が低下する主な原因となる不利な環境下での話者照合について追加で調査を行う必要がある．

謝辞 この成果は，国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）の委託業務（JPNP20006）の結果得られたものです．

参考文献

[1] Snyder, D., Garcia-Romero, D., Povey, D. and Khudanpur, S.: Deep neural network embeddings for text-independent speaker verification., *Interspeech*, pp. 999–1003 (2017).

[2] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. and Khudanpur, S.: X-vectors: Robust dnn embeddings for speaker recognition, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 5329–5333 (2018).

[3] Tang, Y., Ding, G., Huang, J., He, X. and Zhou, B.: Deep speaker embedding learning with multi-level pooling for text-independent speaker verification, *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 6116–6120 (2019).

[4] You, L., Guo, W., Dai, L. and Du, J.: Multi-Task learning with high-order statistics for X-vector based text-independent speaker verification, *arXiv preprint arXiv:1903.12058* (2019).

[5] Kanagasundaram, A., Sridharan, S., Ganapathy, S., Singh, P. and Fookes, C.: A study of x-vector based speaker recognition on short utterances, *Proceedings of the 20th Annual Conference of the International Speech Communication Association, INTERSPEECH 2019. Vol. 2019-September.*, ISCA (International Speech Communication Association), pp. 2943–2947 (2019).

[6] Li, X., Zhong, J., Yu, J., Hu, S., Wu, X., Liu, X. and Meng, H.: Bayesian x-vector: Bayesian neural network based x-vector system for speaker verification, *arXiv preprint arXiv:2004.04014* (2020).

[7] Hong, Q.-B., Wu, C.-H., Wang, H.-M. and Huang, C.-L.: Statistics pooling time delay neural network based on X-vector for speaker verification, *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 6849–6853 (2020).

[8] Xu, L., Das, R. K., Yılmaz, E., Yang, J. and Li, H.: Generative x-vectors for text-independent speaker verification, *2018 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, pp. 1014–1020 (2018).

[9] Li, L., Tang, Z., Shi, Y. and Wang, D.: Gaussian-constrained training for speaker verification, *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 6036–6040 (2019).

[10] Xu, L., Ren, B., Zhang, G. and Yang, J.: Linear transformation on x-vector for text-independent speaker verification, *Electronics Letters*, Vol. 55, No. 15, pp. 864–866 (2019).

[11] Yang, Y., Wang, S., Sun, M., Qian, Y. and Yu, K.: Generative adversarial networks based x-vector augmentation for robust probabilistic linear discriminant analysis in speaker verification, *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, IEEE, pp. 205–209 (2018).

[12] Wu, Z., Wang, S., Qian, Y. and Yu, K.: Data Augmentation Using Variational Autoencoder for Embedding Based Speaker Verification., *INTERSPEECH*, pp. 1163–1167 (2019).

[13] Nidadavolu, P. S., Kataria, S., Villalba, J., Garcia-Perera, P. and Dehak, N.: Unsupervised feature enhancement for speaker verification, *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 7599–7603 (2020).

[14] Kataria, S., Nidadavolu, P. S., Villalba, J., Chen, N., Garcia-Perera, P. and Dehak, N.: Feature enhancement with deep feature losses for speaker verification, *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 7584–7588 (2020).

[15] Susumu Saito, Yuta Ide, T. N. T. O.: VocalTurk: Exploring feasibility of crowdsourced speaker identification, *The 22th Annual Conference of the International Speech Communication Association (INTERSPEECH2021)*, pp. XXXX–XXXX (2021).

[16] Parson, J., Braga, D., Tjalve, M. and Oh, J.: Evaluating voice quality and speech synthesis using crowdsourcing, *International Conference on Text, Speech and Dialogue*, Springer, pp. 233–240 (2013).

[17] Cambre, J., Colnago, J., Maddock, J., Tsai, J. and Kaye, J.: Choice of voices: A large-scale evaluation of text-to-speech voice quality for long-form content, *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13 (2020).

[18] Tarasov, A., Delany, S. J. and Cullen, C.: Using crowdsourcing for labelling emotional speech assets, *W3C workshop on Emotion ML* (2010).

[19] Smith, J., Tsiartas, A., Wagner, V., Shriberg, E. and Bassiou, N.: Crowdsourcing emotional speech, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 5139–5143 (2018).

[20] Gallardo, L. F., Jimenez, R. Z. and Möller, S.: Perceptual Ratings of Voice Likability Collected Through In-Lab Listening Tests vs. Mobile-Based Crowdsourcing., *Interspeech*, pp. 2233–2237 (2017).

[21] Hovy, D., Berg-Kirkpatrick, T., Vaswani, A. and Hovy, E.: Learning whom to trust with MACE, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1120–1130 (2013).

[22] Paun, S., Carpenter, B., Chamberlain, J., Hovy, D., Kruschwitz, U. and Poesio, M.: Comparing Bayesian Models of Annotation, *Transactions of the Association for Computational Linguistics*, Vol. 6, pp. 571–585 (2018).

[23] Martín-Morató, I., Harju, M. and Mesaros, A.: Crowdsourcing strong labels for sound event detection, *arXiv preprint arXiv:2107.12089* (2021).

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

[25] Prince, S. J. and Elder, J. H.: Probabilistic linear discriminant analysis for inferences about identity, *2007 IEEE 11th International Conference on Computer Vision*, IEEE, pp. 1–8 (2007).

[26] Snyder, D., Chen, G. and Povey, D.: Musan: A music, speech, and noise corpus, *arXiv preprint arXiv:1510.08484* (2015).