話者照合のための回り込みを考慮した非線形帯域拡張法と通信音声による評価 *

☆上西遼大, 塩田さやか, △貴家仁志 (首都大東京)

1 はじめに

近年, 声を用いた生体認証技術である話者照合の実用化が進んできている. 特に電話回線を通した使用が見込まれているが, 通信を介した音声は低いサンプリング周波数や帯域制限によって音声の明瞭性や話者性が大幅に低下してしまい認証システムの性能にも大きな影響を与えてしまうことが知られている. この問題に対して, 帯域拡張という広帯域成分を復元・生成する技術を用いることで, 音声の明瞭性や話者性が改善されることが報告されている [1]. これまでに帯域拡張法の一つとして非学習型の非線形帯域拡張法が提案され, 軽い計算量ながら話者照合率が改善することが報告された [2]. しかしその報告ではクリーンな音声での評価のみしか行われていなかった. そこで本稿では電話音声に対して非線形帯域拡張法を適応し i-vector に基づく話者照合 [3] を行い, また客観評価による考察についても報告する.

2 非線形帯域拡張法と通信音声

2.1 回り込みを考慮した非線形帯域拡張法 [4]

非線形帯域拡張法はアップサンプリングされた音声に対して以下の式の非線形関数をかけることで高周波成分を生成する手法である. しかし, 低周波成分にも信号が回り込むため, さらに回り込みを考慮した非線形帯域拡張法が提案されている. 具体的なフローとしては図 1 に示す通りで非線形関数の前とリミッターの後に Filter(A), (B) をかけることで回り込みの影響を緩和している. Filter(B) にはハイパスフィルタ (HPF) やバンドパスフィルタ (BPF) を想定しており, 回り込みを取り除くことで $y_{up}[n]$ との足し合わせの際に元の音声を傷つけないためノイズが低減されると期待できる.

2.2 通信音声

通信を介した音声は様々なパターンがあるが本稿では固定電話の音声について考える. 本稿では入力音声に対してを ITU-T 勧告 G.712[5] に基づくフィルタを用い帯域制限をかけたうえでダウンサンプリングを行う. 次に ITU-T 勧告 G.711[6] によって策定された μ-law 方式による符号化を用いることで固定電話を介した音声を摸擬した.

3 実験

i-vector に基づく話者照合システムにおいて提案法が通信音声に有効であるかを評価した.

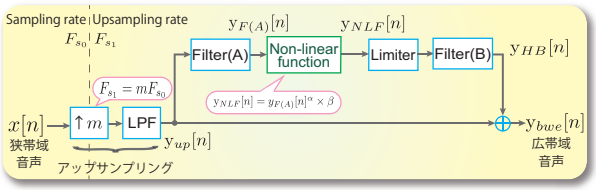

図 1 非線形帯域拡張のフロー

3.1 実験条件

話者照合システムの主な構築条件を示す. i-vector を推定するために必要となる UBM, TV 行列の学習には JNAS データベースから女性話者の音声 23657 文章を用いた. GMM の混合数は 1024, i-vector の次元数は 400 次元である. 登録データとしては VLD データベース [7] から女性話者 17 名の音声を用いている. JNAS データベースはサンプリング周波数が 16kHz だが, VLD データベースはサンプリング周波数が 48kHz であるため 16kHz にダウンサンプリングしたものを 16kHz の原音声として扱う. この原音声のうち, 70 文章 ×17 名を特定話者モデルの学習データ, 30 文章 ×17 名をテストデータとした. フレーム長は 25 ms, フレームシフトは 10 ms, 音響的特徴量は MFCC19 次元+Δ+ΔΔ とした. 図 1 の非線形関数に用いたパラメータは $\alpha = 2, \beta = 100000$ となっており, 全手法で共通したものを用いた. また Filter(A) はどちらもハイパスフィルタとし阻止域周波数 80Hz とした. 話者照合システムの評価には等価エラー率 (Equal Error Rate; EER), 客観評価尺度として PESQ 及び RMS-LSD[8] を用いた. PESQ は値が高いほど自然性が高く, RMS-LSD は値が小さいほど比較音声との誤差が少ないことを示す. 本実験では比較音声は全てサンプリング周波数 16kHz の原音声とする. また実験において μ は圧縮効率を表しており, 値が小さいほど強く圧縮されていることを示す.

3.2 比較条件

本研究で比較する主な手法は次の 3 手法である: ①入力音声を 8kHz から 16kHz にアップサンプリング. ②入力音声に非線形帯域拡張を適用. ③8kHz のままのもの. これらの比較手法①~③に対して入力する音声がクリーンな音声である場合をそれぞれ (A), (B), (C), 圧縮パラメータを $\mu = 255$ として電話音声を模擬したものを (D), (E), (F), $\mu = 127$ の場合を (G), (H), (I), $\mu = 63$ の場合を (J), (K), (L) とした. 全ての手法において入力音声は 16kHz の原音声から 8kHz にダウンサンプリングされたものを用いている.

---

*Nonlinear artificial bandwidth extension considering aliasing artifacts for speaker verification and its evaluation to communication signals by KAMINISHI Ryota, SHIOTA Sayaka, KIYA Hitoshi (Tokyo Metropolitan University)

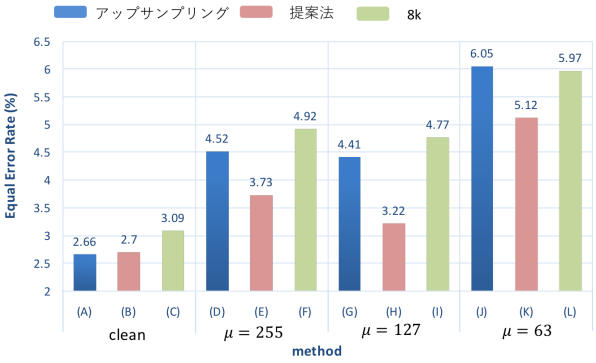

図 2　話者照合結果

3.3　実験結果

図 2 に手法ごとの EER を示す．まず，クリーン音声における (A) アップサンプリング，(B) 提案法，(C)8k を比較すると，(C) よりもサンプリング周波数をあげた (A)，(B) の方が EER が低くなっていることがわかる．ここで (A) が (B) よりも EER が低いのは (A) はノイズも含まれず重要な情報である低周波成分が汚れていないためモデル化がうまくできたことが要因であると考えられる．ただし，(B) と比べて有意な差ではない．次に $\mu = 255$ のときの (D)，(E)，(F) について比較してみる．この 3 つの手法の中で提案法を用いた (E) が一番照合性能が良いことがわかる．$\mu = 127$，63 も同様の傾向が得られた．圧縮がかかりノイズを含む音声においては非線形帯域拡張法が有効であることがわかる．これは提案法により生成した高周波数成分がノイズの影響を受けていても話者性を表現できているからだと考えられる．$\mu = 255$ 及び $\mu = 127$ の結果を比較するときつい圧縮がかかる $\mu = 127$ の方が 3 手法とも EER が若干低い．8k の結果でも EER が低いことからノイズが含まれていても話者性を表す部分には悪い影響を与えておらず，結果として (G)，(H) の EER が (D)，(E) よりも低くなったと考えられる．

次に客観的評価尺度である PESQ 及び RMS-LSD を用いた結果を図 3，4 に箱ひげグラフで示す．箱の上辺と底辺は全結果の四分値範囲を箱の中の線はデータの中央値であり，具体的な数値は手法名の上に載せてある．箱の上下に伸びる線は全データの最大値と最小値を示す．図 3 において提案法 (E)，(H)，(K) とアップサンプリング (D)，(G)，(J) を比較するとアップサンプリングのみの方はどの条件においても自然性が高くなっている．しかし，図 4 において提案法 (E)，(H)，(K) とアップサンプリング (D)，(G)，(J) を比較すると提案法の手法はどの条件においてもより誤差が少なくなっている．提案法は自然性を上げることを目的としていないため，自然性を評価する PESQ の値が改善されることは約束されないが，スペクトルの誤差としてはアップサンプリングよりも非常に低くなっている．これは提案法により高周波成分が生成されているためだと考えられる．圧縮によるノイズが強

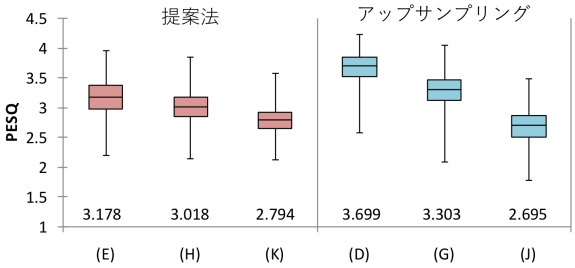

図 3　PESQ による評価

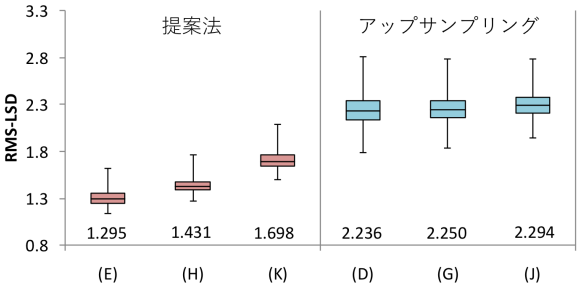

図 4　RMS-LSD による評価

くなった場合 PESQ 及び RMS-LSD どちらも値が悪くなるものの同様の傾向であることから提案法が有効であることが確認できた．

4　おわりに

本稿では回り込みを考慮した非線形帯域拡張法が通信音声に適応した場合を評価した．実験結果より，サンプリング周波数 8kHz をそのまま用いるよりも非線形帯域拡張を用いる方が照合性能が改善されることを確認した．このことより通信音声においても非線形帯域拡張法が有効であることがわかった．今後の課題は異なる圧縮方式を適応する場合や PLDA の適用などがあげられる．

謝辞 本研究の一部は科学研究費基盤 (B)2628006 による．

参考文献

[1] P. Robert et al., In Proc. ICASSP, pp.3699–3703, 2014.
[2] 中西涼介ら, 日本音響学会春季大会, pp.131–134, 2017.
[3] N. Dehak et al., IEEE Trans., vol.19, pp.788–798, 2010.
[4] 上西遼大ら, 電子情報通信学会 音声研究会, pp.29–32, 2017.
[5] ITU-T Recommendation G.712, 1987.
[6] ITU-T Recommendation G.711, 1987.
[7] S. Shiota et al., In Proc. Interspeech, pp.239–243, 2015.
[8] P. Jax et al., Signal Processing, vol.83, no.8, pp.1707–1719, 2003.