6U-3

若年話者判別法の音響特徴に対する聴覚フィルタバンクの導入

宮森 翔子　西村 竜一　岡本 恵里香　河原 英紀　入野 俊夫
和歌山大学 システム工学部

1 はじめに

本研究では，特に対話システムへの応用を視野に入れた，発話情報による子ども利用者の判別法を提案する．現在，たばこの購入に際した年齢確認 [1] や，自動販売機における販売促進のために，生体情報による年齢層の自動判別が応用されている [2]．対話システムにおいても，利用者の年齢層を判別し，反応を変化させることで，より利用者の満足度の高い対応を提供できる．

これまで本研究では，音響特徴量に MFCC(Mel Frequency Cepstrum Coefficient) を用いた判別法について検討してきた．しかし，変声期にあたる 10 代後半の話者に対する判別性能が低いことが問題となっていた [3]．また，和田らによる MAP 適応した GMM や MLLR 変換行列を用いた手法の提案では，子ども話者に対する検討が不足している [4]．さらに，従来の特徴量は，音声認識や話者認識で一般的に使用される特徴量であり，年代推定や子ども判別に必ずしも適していない可能性がある．

今回，従来法における問題を解決するために，新たな特徴量として，ガンマチャープ聴覚フィルタバンク (GCFB:Gammachirp Filterbank)[5] による特徴量抽出を行った．先行研究により，GCFB による声道長比正規化が音声モーフィングに有効であるという結果が得られている [6]．ここで推定される，声道長は身長や体重に相関がある [7] ため，声道長の情報は大人と子どもを判別するために有効になり得る．そのため，GCFB による特徴量の導入と判別実験を行った．

2 HMM 判別器を用いた比較評価実験

従来法との比較のために，HMM(Hidden Markov Model) を判別器として用いた実験を行った．評価を行うに当たり，本研究では，大人と子どもの境界となる年齢を年齢閾値として設定した．例えば，年齢閾値 15 歳の場合は，15 歳未満の発話者を子ども，15 歳以上の発話者を大人とみなすことになる．実験では，年齢閾値を 9 歳から 20 歳まで 1 歳ごとに変化させて，各年齢閾値ごとにモデルの学習および判別を行った．

学習段階では，まず，発話者の年齢，性別および年齢閾値に基づき，学習用発話をクラス分けする．クラスは大人女性，大人男性，子どもの 3 クラスとした．次に，各クラスについて収集発話から特徴量を抽出し，発話単位でモデルを作成する．HMM の状態数は 3，各状態における GMM の混合数は 128 である．判別段階では，評価用発話に対して最も尤度の高いクラスを結果とする．

実験で使用した発話は，ウェブ上で収集した 2360 発話である．これらは，対話システムの利用環境を想定して整備された発話であり，人手によって録音状態を確認した [8]．評価は，10 分割交差検定により行い，学習用発話に評価用発話の話者が含まれない状態 (話者オープン) とした．モデルの構築には HTK3.4.1 [9]，判別デコーダには Julius4.1.4 [10] を使用した．

2.1 実験に使用した特徴量

従来法として，MFCC と Δ および ΔPower の 25 次元からなる特徴量を使用した．これは音声認識や話者認識などで一般的に使用される特徴量である．

一方，提案法では，ガンマチャープ聴覚フィルタバンク (GCFB) を用いて抽出した特徴量を使用した．GCFB は，聴覚末梢系の周波数分析機能をより模擬するためのフィルタバンクである．これは，人間の周波数分析特性を心理物理的に測定してモデル化した聴覚フィルタ [11] を，疑似対数周波数軸上に並べたものである．さらに今回は，末梢系の動的なスペクトル分析をより高精度に近似する手法を使用している [5]．これにより，入力音圧に依存したフィルタ形状変化や，圧縮特性の模擬が実現され，2 音抑圧等の非線形特性も説明できている．

今回，GCFB を使用した特徴量として，2 種類の特徴量を用意した．1 つ目は，25ch+Δ は，25ch 分のフィルタ出力とその変化量 (Δ) を用いたものである．もう一方は，GCMC(Gammachirp Modulation Coefficients) では，GCFB の出力の対数を求めた後，離散コサイン変換を適用した 13 次元の係数に加え，その各係数の時間変動の低周波成分 (2〜16Hz) を加えた合計 25 次元を特徴量とした [12].

3 評価実験結果

F 値による比較を行った．F 値は，情報検索システムの性能を表す総合的な評価尺度であり，適合率及び再現率の調和平均で求めることができる．F 値が 1 に近づくほど，システムの性能はよいとされる．

図 1 に F 値による大人・子ども判別結果の比較を示す．図の横軸は年齢閾値である．図中の緑線は 25ch+Δ，青線は従来法の MFCC，赤線は GCMC による結果を示す．GCFB を用いた特徴量のうちで比較すると，年齢閾値 20 歳を除き，25ch+Δ より GCMC が高い値を示している．MFCC の F 値は，年齢閾値 14 歳から 15 歳に 0.065，18 歳から 19 歳にかけて 0.09 下がるなど，年齢閾値の変化に対して安定ではない．比較して，GCMC では年齢閾値 15 歳以降における年齢閾値 1 歳ごとの変化が平均 0.03 と緩やかであり，比較的安定だといえる．

さらに，図 2，3 に，年齢閾値 18 歳における判別結果のクラスを割合で示した．グラフの横軸は年齢，縦軸は判別結果として得られたクラスの割合を示す．各グラフの横軸の右端では，20 歳代と 30 歳以降の結果をまとめて示している．図中の緑の部分は子ども，赤は大人女性，青は大人男性クラスへと判別されたということを示す．このグラフは年齢閾値 18 歳の結果であるため，18

Development of Acoustic Features Based on Gammachirp Filterbank for a Child Speaker Identification, Shoko MIYAMORI, Ryuichi NISIMURA, Erika OKAMOTO, Hideki KAWAHARA, Toshio IRINO (Faculty of Systems Engineering, Wakayama University)

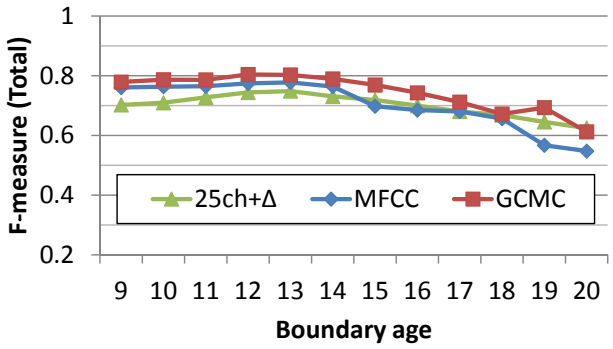

図 1: 評価実験結果（F 値）

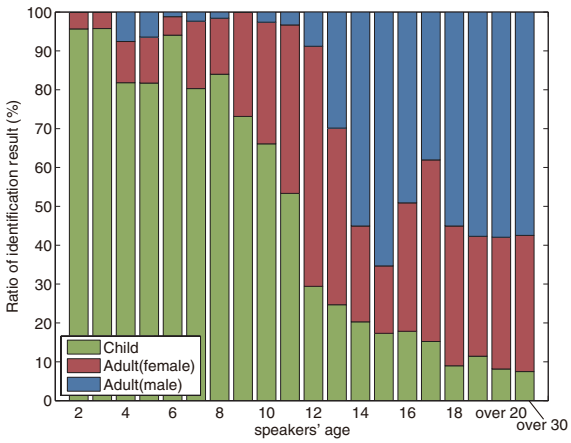

図 2: MFCC による評価実験結果 (年齢閾値 18 歳)

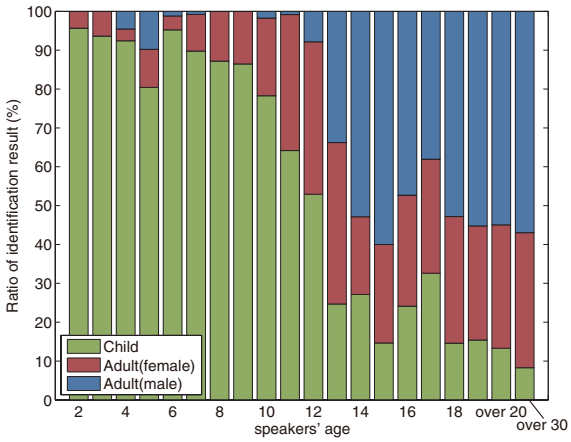

図 3: GCMC による評価実験結果 (年齢閾値 18 歳)

歳未満を全て正しく判別した場合，グラフの色は 17 歳以下で全て緑となる．GCMC では MFCC に比べ，8 歳から 12 歳にかけての判別性能が平均で 12%と特に向上している．また，13 歳から 17 歳にかけても，5%の向上がみられた．

これらの結果から，GCMC は全体的に MFCC より判別性能の向上を確認できた．また，GCFB を用いた特徴量のうちでも，GCMC の性能が高いことが分かった．今回，GCMC で用いた時間変動の低周波成分は，音声に特有な特徴を表現したものである．これを抽出したことにより，環境に依存する要因を抑制できた可能性がある．

4 まとめ

本研究では，対話システムのための若年話者判別について，聴覚フィルタバンクにより抽出した特徴量を導入した．従来の特徴量である MFCC と，3 クラス HMM による判別実験において性能の比較を行った．その結果，GCMC での判別において，従来法より F 値が平均 0.6 上回った．これは，GCMC に含まれる低い周波数の変調成分の効果だと考えられる．

謝辞 本研究の一部は、和歌山大学平成 23 年度独創的研究支援プロジェクト及び科学研究費補助金の支援を受けた。

参考文献

[1] 株式会社フジタカ, ”成人識別装置「こどもチェックシステム」”, http://www.fujitaka.com/kao/, 2007.

[2] インテル株式会社, ”効率的な運用と宣伝機能を備える将来の自動販売機コンセプトを公開”, http://www.intel.com/jp/intel/pr/press2010/100512a.htm, 2010.

[3] 宮森 他, ”ちょっとした一言の音声認識による子ども利用者判別法の検討,” FIT2010 第 9 回情報科学技術フォーラム, pp.469-472, 2010.

[4] Toshiya Wada, et al., ”Investigations of Features and Estimators for Speech-based Age Estimation”, Proc. APSIPA, 2010.

[5] Irino, T. and Patterson, R.D., ”A dynamic compressive gammachirp auditory filterbank,” IEEE Trans. Audio, Speech, and Language Process., vol. 14, no. 6, pp.2222–2232, 2006.

[6] Okamoto, E., et al., ”Auditory Filterbank Improves Voice Morphing,” Proc. Interspeech 2011, pp.2517 - 2520, 2011.

[7] W.T. Fitch, J. Giedd, ”Morphology and development of the human vocal tract: a study using magnetic resonance imaging,” J. Acoust. Soc. Am., vol. 106, pp.1511–1522, 1999.

[8] 栗原 他, ”音声ウェブシステムを用いて収集した実環境子供発話に関する調査”, FIT2010 第 9 回情報科学技術フォーラム講演論文集, pp.229 - 230, 2010.

[9] Young, S.J., et al., ”The HTK book version 3.4”, Cambridge University Engineering Department, Cambridge, UK, 2006.

[10] Lee, A., et al., ”Julius — An Open Source Real-Time Large Vocabulary Recognition Engine”, Proc. Eurospeech 2001, pp.1691–1694, 2001.

[11] 入野俊夫, ”聴覚フィルタの測定と定式化について,” 日本音響学会聴覚研究会資料, H-2009-73, Vol.39, No.6, pp.413-418, 2009.

[12] HariKrishna, M., Marco, M., ”A Level-dependent Auditory Filter-bank for Speech Recognition in Reverberant Environments,” Proc. Interspeech 2011, pp.685–688, 2011.