クロストーク音声認識における同時発話認識率の調査

9 9 2 0 1 7

計算機工学講座　池原研究室　岡本 一輝

1　はじめに

クロストーク音声認識は技術的に困難な課題であり，従来，研究例が少ないが，現実の音声認識では重要な技術の一つである．最近，男女 2 話者の単独同時発話を対象に，現状の技術を用いた認識率の実験的評価が行われている [1] が，実験対象とする単語数が多いこともあって，低い認識率にとどまっている．また，実験では，片側音声を対象とした単独認識率のみが評価されており，両側音声の同時発話認識率は不明である．そこで，本研究では，認識対象単語数と認識率の関係を調べるため，10 単語を対象とした認識実験を行い，同時発話認識率についても評価する．

2　評価実験

音声データベースとして，ATR 単語発話データベース Aset の男女各 2 名を使用する．奇数番音声を学習データ，偶数番音声を認識データに使用する．認識データから 4 モーラで，発話時間がほぼ同じ語をランダムに 10 単語ずつ抽出する．それぞれを重ね合わせて作成した 100 単語のクロストーク音声を作成し，実験データとする．表 1 に実験に使用した単語を示す．また図 1 にクロストーク音声認識の手順を示す．

表 1: 実験に使用した単語

男性話者	女性話者
悪質 (akusitsu)	足元 (asimoto)
聞こえる (kikoeru)	可愛い (kawaii)
加える (kuwaeru)	勤勉 (kiNbeN)
失恋 (sitsureN)	細々(komagoma)
優れる (sugureru)	すまない (sumanai)
そのうち (sonouchi)	対策 (taisaku)
中毒 (chuudoku)	手拭い (tenugui)
内容 (naiyou)	天才 (teNsai)
暴力 (bouryoku)	滅ぼす (horobosu)
わざわざ (wazawaza)	欲張る (yokubaru)

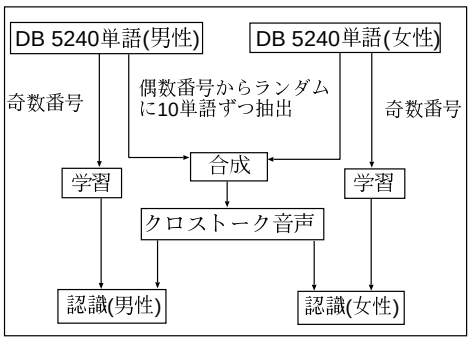

図 1: クロストーク音声認識の手順

2.1　実験条件

本実験では HTK[2] を使用して実験を行う．実験環境を表 2 にまとめる．特徴パラメータに FBANK と MFCC を使用する．また音素 HMM の共分散行列には Diagonal-covariance 及び，Full-covariance を使用し，単独認識率，及び同時発話認識率を調査する．なお，単独認識率は，男性話者または女性話者を単独で認識した場合の認識率である．また同時発話認識率とは男性話者と女性話者を同時に認識できた場合の認識率である．

2.2　実験結果

表 3 に MFCC 及び FBANK の認識結果を示す．特徴パラメータに FBANK，共分散行列に Full-covariance を用いた手法が最も認識率が高く，同時発話認識率は 54%であった．

表 2: 実験条件

	MFCC	FBANK
基本周波数	16kHz	
分析窓	Hamming 窓	
分析窓長	25ms	
フレーム周期	10ms	
音響モデル	3 ループ 4 状態・半連続分布型	
stream 数	2	
混合分布数	128mixture	
特徴パラメータ	MFCC 12 次 +対数パワー (計 13 次)	FBANK 24 次 +対数パワー (計 25 次)

表 3: 実験結果

	単独認識率 (男性話者)	単独認識率 (女性話者)	同時発話認識率
MFCC Diagonal-covariance	62%	79%	46%
MFCC Full-covariance	67%	79%	51%
FBANK Diagonal-covariance	60%	54%	31%
FBANK Full-covariance	72%	78%	54%

2.3　人間による聴覚実験

計算機による認識率と比較するため，実験に使用したクロストーク音声に対して，人間による聴覚実験を行い，同時発話認識率を求めた．なお，被験者は男性 3 名，女性 1 名である．表 4 に認識結果を示す．同時発話認識率の平均が 77%であった．

表 4: 実験結果 (人間による聴覚実験)

	単独認識率 (男性話者)	単独認識率 (女性話者)	同時発話認識率
人間による認識率	85%	91%	77%

3　考察

計算機による実験結果と人間による聴覚実験結果と比較すると，最も結果の良い FBANK Full-covariance においても計算機による認識率が低い．また人間による認識での誤り率の改善は単独認識率も同時認識率も約 50%程度である．計算機では撥音と濁音が入る音声で誤認識が起こりやすく，人間では実験で使用した音声データの子音の発音が弱い音声で誤認識が起こりやすい傾向がある．具体例を表 5 に示す．

表 5: 認識結果例

人間で聞き取れたが、計算機で認識できなかった例		
正解 (男性/女性)	加える (kuwaeru)	天才 (teNsai)
計算機の結果	加える (○)	手拭い (tenugui)
聴覚実験結果	加える (○)	天才 (○)
計算機で認識できたが、人間で聞き取れなかった例		
正解 (男性/女性)	内容 (naiyou)	対策 (taisaku)
計算機の結果	内容 (○)	対策 (○)
聴覚実験結果	概要 (gaiyou)	対策 (○)

4　おわりに

男女 2 話者が別々の孤立単語を同時に発声する状況を想定し，4 つの手法を用いて認識率を調査した．その結果，FBANK Full-covariance を用いた手法が最も認識率が高い結果となった．しかし，人間による聴覚実験と比べると認識率は非常に低い．今後は，今回の研究で得られた結果を基礎データとして，雑音対策に使用されている手法の適用を行い認識率を向上させていく予定である．

参考文献

[1] 安田:クロストーク音声認識，鳥取大学工学部知能情工学科卒業論文 (2003).

[2] Seve Young, et al.:HTK Ver3.2 Reference manual, Cambridge University(2002).