音声の構造的表象を用いた加算性雑音音声認識*

◎村上隆夫, 丸山和孝, 峯松信明, 広瀬啓吉 (東大)

1　はじめに

従来の音声認識技術は音声の音響的実体を対象としてきたが、この実体には話者の声道形状の特性、マイクロフォンの特性などの非言語的特徴によって不可避的に歪む。このため、不特定話者音響モデルなどの「集める」ことによる解決策には限界があると考えられる。近年、上記の非言語的特徴を表現する線形変換性歪み・乗算性歪みを原理的に持たない音響的普遍構造が提案された [1, 2]。これは、構造音韻論の物理実装として、あるいは音声ゲシュタルトとして解釈される音声の構造的表象である [1, 2]。[3] では、加算性雑音が母音のスペクトル高域成分に多く含まれる話者性の情報 [4] を消失させることを示した。本稿では、構造を用いた雑音下の日本語母音系列の認識実験について、従来手法との比較実験を含めて報告する。

2　雑音下の日本語母音系列音声認識

2.1　構造を用いた日本語母音系列音声認識の枠組み

本稿では日本語母音系列を認識タスクとする。各母音は孤立発声され、一回ずつ出現する（語彙サイズ：${}_5P_5 = 120$）。このような認識タスクを構造のみ用いて認識する枠組みは以下のとおりである。まず、入力音声から距離行列を求めることで音声を構造化する。この際、構造サイズは調音努力を表し [5]、また、雑音下では構造サイズが小さくなる [3] ので、これを正規化する。距離行列のうち意味を持つ成分は上三角成分であるので、これをベクトルとして並べた「構造ベクトル」（10 次元）を特徴量として用いる。次に、認識器に持たせる構造モデルは以下のようにして得る。複数の/a/-/i/-/u/-/e/-/o/構造ベクトルから 10 次元ガウス分布を求め、これを「構造統計モデル」として使用する。他の 119 個のモデルは/a/-/i/-/u/-/e/-/o/モデルの要素を交換することで得られる。構造の音響的照合は、入力構造ベクトルと各構造統計モデルのマハラノビス距離を求めることで行なう。構造を用いた日本語母音系列音声認識の枠組みを Fig. 1 に示す。

2.2　クリーンな構造統計モデルを用いた認識実験

[3] では、加算性雑音による構造の歪みについて報告した。雑音を重畳させた音声をクリーンな構造統計モデルを用いて認識する場合、入力音声の構造の歪みが原因で認識性能が低下することが予想される。これをより詳細に調べるため、以下の実験を行なった。

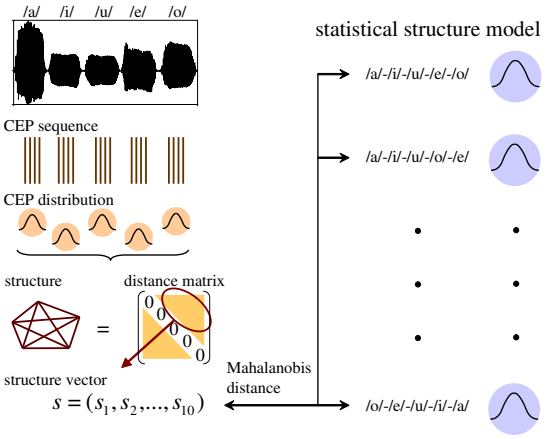

Fig. 1　構造を用いた日本語母音系列音声認識

音声データとしては、8 名話者（男性 4 名、女性 4 名）が 5 母音を 5 回ずつ孤立発声したものを用いた。ここから各話者毎に 3,125（$= 5^5$）個の/a/-/i/-/u/-/e/-/o/の音声を得た。その各々に SNR=∞（clean）, 20, 10, 0[dB] の白色雑音を重畳し、LPF（カットオフ周波数：2kHz）を施した後、入力構造ベクトル（計 $8 \times 5^5 = 25,000$ 個）を得た。このとき、音声事象分布の推定は ML or MAP の両方を試みた。MAP の重みは $n = 10, 1, 0.1, 0.01$ のうち最適なものを選んだ。また、雑音による影響を軽減するため、LPF 後に SS（$\alpha = 2.0$、$\beta = 0.5$）を行なう場合も試みた。その際、雑音パワースペクトルの推定には 300ms の白色雑音区間を用いた。構造統計モデルは、評価話者を除く 7 名によるクリーン音声から得た計 21,875（$= 7 \times 5^5$）個の/a/-/i/-/u/-/e/-/o/構造ベクトルを用いて学習させた。音響的条件を Table 1 に示す。

実験結果は Table 2 のとおりである。雑音を重畳することで認識性能が劣化していることが分かる。低 SNR のときに MAP 推定による効果が見られなくなったのは、事前知識の推定をクリーンな環境で行なっているため、雑音下の入力音声との間でミスマッチが生じたことが原因と考えられる。

2.3　雑音下の構造統計モデルを用いた認識実験

[6] では、一人の男性話者音声によって学習された構造統計モデルを用いて、日本語母音系列のクリーン音声に対する 100%の認識性能を実現している（この際、カットオフ周波数が 2kHz の LPF を用いている）。構造統計モデルの学習に必要な話者が一人で十分ならば、極めて高品質な音声合成器を用いて、評価

*Automatic recognition of speech in noise using the structural representation of speech
By Takao Murakami, Kazutaka Maruyama, Nobuaki Minematsu and Keikichi Hirose (University of Tokyo)

Table 1　音響的条件

サンプリング	16bit / 16kHz
窓	窓長 25msec、シフト長 10msec
パラメータ	MCEP（α=0.55）（1〜12 次元）
分布推定方法	ML or MAP

Table 2　クリーンな構造統計モデルを用いた実験結果

	w/o SS		SS	
SNR	ML	MAP	ML	MAP
∞	82.9%	99.9%	-	-
20[dB]	55.0%	98.5%	68.7%	99.9%
10[dB]	38.5%	39.5%	57.0%	52.8%
0[dB]	12.7%	12.4%	18.2%	13.1%

音声の雑音環境と合致する音声を合成し、それを基に構造統計モデル（及び事前知識）をオンラインで学習させることも可能と考えられる。少なくとも人間は完璧な合成器を持っており、このことは構造に基づく音声知覚の運動理論 [7] と解釈することができる。

ここでは、雑音下の構造統計モデルの性能を調べるため、評価音声の SNR が既知との仮定のもと、男性話者 1 名の音声データを用いて雑音下の構造統計モデル（及び事前知識）を学習させる実験を行なった。男性話者が 5 母音を 35 回発声したデータを 7 つのグループに分け、各グループ毎に 3,125（$= 5^5$）個の/a/-/i/-/u/-/e/-/o/の音声を得た。その各々に、評価音声と同じ SNR となるような白色雑音を重畳した。ここから、計 21,875（$= 7 \times 5^5$）個の/a/-/i/-/u/-/e/-/o/構造ベクトルを求め、構造統計モデルの学習に用いた。音響的条件は Table 1 と同じである。但し、白色雑音を重畳することで、スペクトル高域成分を揃えることができる [3] ので、LPF を用いない場合（full band）についても試みた。また、SS は行なっていない。

結果を Table 3 に示す。Table 2 よりはるかに良い認識性能が得られている。これは、入力構造と構造統計モデルとの間で雑音環境のミスマッチが無くなったことが原因と見られる。また、全帯域を用いた場合においては、雑音環境下の方が高い認識率が得られ、低 SNR では LPF を施した場合よりも良い性能が得られている。これは、白色雑音を重畳することで、フォルマントの情報を保ちつつ、スペクトル高域成分を揃えることができたためと思われる。但し、雑音レベルを非常に大きくする（SNR = 0[dB]）と認識性能が劣化してしまう。これは音声が雑音に埋もれて音韻差異が不明瞭になるためと考えられる。

2.4　従来手法との比較実験

比較のため、SS（$\alpha = 2.0$、$\beta = 0.5$）を用いた従来手法による認識実験も行なった。雑音パワースペクトルの推定には 300ms の白色雑音区間を用いた。音響

Table 3　雑音下の構造統計モデルを用いた実験結果

	full band		2kHz	
SNR	ML	MAP	ML	MAP
∞	24.7%	70.3%	86.8%	100.0%
20[dB]	73.9%	92.9%	67.9%	99.8%
10[dB]	77.4%	99.1%	68.1%	86.7%
0[dB]	73.9%	87.0%	71.1%	85.1%

Table 4　従来手法との性能比較

SNR	HMM(260)	HMM(4,130)	Proposed(1)
∞	100.0%	100.0%	100.0%
20[dB]	100.0%	98.8%	99.8%
10[dB]	94.3%	97.2%	99.1%
0[dB]	83.0%	86.8%	87.0%

モデルは、学習話者 4,130 名の混合共有 HMM、学習話者 260 名の状態共有 HMM の 2 通りの不特定話者モデルを用いた。特徴量としては、全帯域の MFCC（1〜12 次元）、Δ MFCC（1〜12 次元）、及び Δ E を用い（計 25 次元）、CMN による話者・環境の正規化も行なった。言語的制約としては、120 単語のみを許容する文脈自由文法を用いた。

実験結果を Table 4 に示す。提案手法の認識性能も合わせて載せている。括弧内の数値は、学習話者数である。低 SNR において、提案手法はいずれの従来手法よりも良い性能を得ていることが分かる。

3　まとめ

構造を用いて、加算性雑音下の日本語母音系列を認識する実験を行なった。その結果、評価音声と同じ SNR の構造統計モデル（及び事前知識）を用いることで、男性話者 1 名の音声で学習された提案手法が、4,130 名の音声で学習された従来の HMM（CMN 及び SS を適用）を上回る結果が得られた。これは、極めて高品質な合成器があれば、オンラインで雑音下の構造統計モデル（及び事前知識）を作成し、高い認識性能を得ることができることを示唆する。今後は、子音を含めた連続音声認識、さらには従来手法との融合を検討していく予定である。

参考文献

[1] N.Minematsu, Proc. ICASSP, pp.899-892 (2005)
[2] 峯松他，信学技報，SP2005-12，pp.1-8 (2005)
[3] 村上他，音講論，1-P-1 (2005-9)
[4] T.Kitamura et al., JASJ(E) Vol.16, No.5 (1995)
[5] N.Minematsu et al, Proc. IWMMS, pp.69-79 (2004)
[6] 村上他，信学技報，SP2005-14，pp.13-18 (2005)
[7] 柏野他，音講論，1-2-10，pp.243-246 (2004-9)