論文 / 著書情報
Article / Book Information

論題(和文)	音響的な認識誤りフレームに対するケプストラム空間拡張の効果
Title(English)	The effect of cepstrum space expansion for acoustically misrecognized frames
著者(和文)	中村 匡伸, 岩野 公司, 古井貞熙
Authors(English)	Masanobu Nakamura, Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会2008年秋季講演論文集, , No. 2-P-17, pp. 395-396
Citation(English)	, , No. 2-P-17, pp. 395-396
発行日 / Pub. date	2008, 9

音響的な認識誤りフレームに対するケプストラム空間拡張の効果*

◎中村 匡伸, 岩野 公司, 古井 貞熙 (東工大)

1 はじめに

話し言葉音声の音響的特徴の分析は，話し言葉音声の認識性能の向上や，音声合成の品質向上に役立つと考えられ，非常に重要である．我々は既に，日本語話し言葉コーパス (以下 CSJ と呼ぶ) に収録されている不特定話者の読み上げ音声と話し言葉音声を用いて各音素のケプストラム特徴量に関する比較を行った．その結果，話し言葉音声では読み上げ音声に比べて全音素間のマハラノビス距離が縮小することにより，音素認識性能が低下していることが明らかになった [1]．また対話音声のような自発性の高い話し言葉音声において，ケプストラム空間の縮小が，認識誤りを引き起こす大きな要因の一つであることが明らかになった [2]．

これまでの我々の研究では，読み上げ音声・話し言葉音声のそれぞれのスタイルについて，音声データ全体を利用して大局的なケプストラム空間の統計量を取り出し，それらを比較することによって分析を行っていた．本研究では，このような大局的な統計量を利用するのではなく，個々のスタイルの音声データに対して，音響的な要因で認識誤りが生じていると考えられる区間（フレーム）を推定し，その区間のケプストラム空間の大きさを他の区間のものと比較することで，局所的なケプストラム空間の縮小に関する分析を行う．また，その縮小による認識性能への影響を分析するために，対象となる区間に対してケプストラム空間を拡張させた音声データを認識に用いた場合の性能について述べる．

2 音声データ

分析には，話し言葉音声として CSJ のテストセットに含まれる学会講演音声 (20 講演，4.0 時間)，読み上げ音声として朗読音声・再読み上げ音声 (20 話者，2.3 時間) を用いた．音声データは 16 kHz でサンプリングされている．実験に際して，まず人手で作成された時間ラベル付きの書き起こしに基づいて音声データを 400 ms 以上の無音区間で区切り，区切られた区間を「発話単位」として定義した．発話単位が 1 秒未満の場合には，後続する発話単位と接続し，1 つの発話単位とみなした．

音響特徴量としては MFCC 12 次元と対数パワー，およびその一次微分，二次微分成分の計 39 次元の音響パラメータを抽出した．分析周期は 10ms, 分析窓幅は 25ms とし，発話単位ごとに CMS 処理を行っている．

本分析では，音響的な認識誤りが生じているフレームの同定を行うために，時間情報を含む音素・単語単位の正解書き起こしと，音声認識結果を必要とする．音声認識に用いる音響モデルには CSJ のテストセットを除いたすべての学会講演音声 (967 講演，234.1 時間) を用いて学習した，left-to-right 型 3 状態の Tri-phone HMM (総状態数 3000，状態あたりの混合数 64) を利用した．言語モデルは CSJ の学会講演音声と模擬講演音声を用いて学習した 30,000 語彙の 2-gram および 逆向き 3-gram を用い，音声認識デコーダは julius(ver4.0) を用いた．

3 音響的誤りフレームの同定

認識誤り要因の同定には，南條らによる分類木を用いたスコア比較による認識誤り原因の分類法 [3] に基づいた，フレーム単位での同定手法を用いた．南條らの提案方法では，認識誤り単語を含む「誤り区間」を設定し，その区間において単語を単位として決定木を用い，誤り原因を分類する．ここで「誤り区間」とは，認識器によって誤りの原因となり得る区間を意味する．本分析では言語モデルとして順向き bigram と逆向き trigram を使っているため，誤り単語を含め前 2 単語と後 1 単語を含む範囲を誤り区間と定義する [3]．誤り区間が重複する場合にはそれらの区間をマージする．以降では簡単のために，例えば音響的な要因による誤りと同定されたフレームのことを「音響的誤りフレーム」と呼ぶ．

Fig. 1 誤り同定分類決定木

1. 認識誤り区間を決定
2. $\{S_a^{(\mathrm{rec})}(t) + S_l^{(\mathrm{rec})}(t) < S_a^{(\mathrm{cor})}(t) + S_l^{(\mathrm{cor})}(t)\}$?
   - YES → サーチエラー
   - NO → 3 へ
3. $\{S_l^{(\mathrm{rec})}(t) > S_l^{(\mathrm{cor})}(t)\} \wedge \{S_a^{(\mathrm{rec})}(t) > S_a^{(\mathrm{cor})}(t)\}$?
   - YES → 音響・言語的要因による誤り
   - NO → 4 へ
4. $\{S_a^{(\mathrm{rec})}(t) > S_a^{(\mathrm{cor})}(t)\}$?
   - YES → 音響的要因による誤り
   - NO → 5 へ
5. 言語的要因による誤り

図 1 に，フレーム単位の誤り原因の分類に用いる決定木を示す．ただし，$S_a^{(\mathrm{cor})}(t)$, $S_l^{(\mathrm{cor})}(t)$, $S_a^{(\mathrm{rec})}(t)$, $S_l^{(\mathrm{rec})}(t)$ をそれぞれ正解単語列における t フレーム目の音響スコアと言語スコア，および認識単語列における t フレーム目の音響スコアと言語スコアとする．誤り区間 W における単語列を $w_1, w_2, \cdots, w_N$，音素列を $q_1, q_2, \cdots, q_M$ とし，w, q の時間長を $l(w), l(q)$ で表す．t 番目のフレームが割り当てられる単語の ID を i_t，音素の ID を j_t，t フレーム目の観測系列を $\mathcal{O}(t)$ とすると，図 1 中の音響スコア $S_a(t)$ は，各音素ごとに求めた音響尤度をフレーム数で平均した値，言語スコア $S_l(t)$ は，各単語ごとに求めた言語尤度をフレーム数で平均した値で，以下のように表される．

$$S_a(t) = \frac{1}{l(q_{j_t})} \sum_{t=\{k|j_k=j_t\}} \log p\big(\mathcal{O}(t) \mid q_{j_t}\big)$$

$$S_l(t) = \frac{\log p(w_{i_t})}{l(w_{i_t})}$$

4 音響的誤りフレームにおける原点からの距離の大きさ

表 1 に，前節の方法で同定された各誤り原因の生起頻度と生起率を示す．表中の SearchErr, ALErr, AcoErr, LngErr, Corr はそれぞれ,「サーチエラーフレーム」,「音響・言語的誤りフレーム」,「音響的誤りフレーム」,「言語的誤りフレーム」，誤り区間に属さない「正解フレーム」を表し，数値は各誤り原因の生起頻度，括弧内は全フレーム数における各誤り原因の生起率 (%) を表す．表 1 より，読み上げ音声に比べて話し言葉音声では音響的誤りフレームの生起率が大きいことが分かる．これはすなわち，話し言葉音

*The effect of cepstrum space expansion for acoustically misrecognized frames, by NAKAMURA, Masanobu, IWANO, Koji, and FURUI, Sadaoki (Tokyo Institute of Technology).

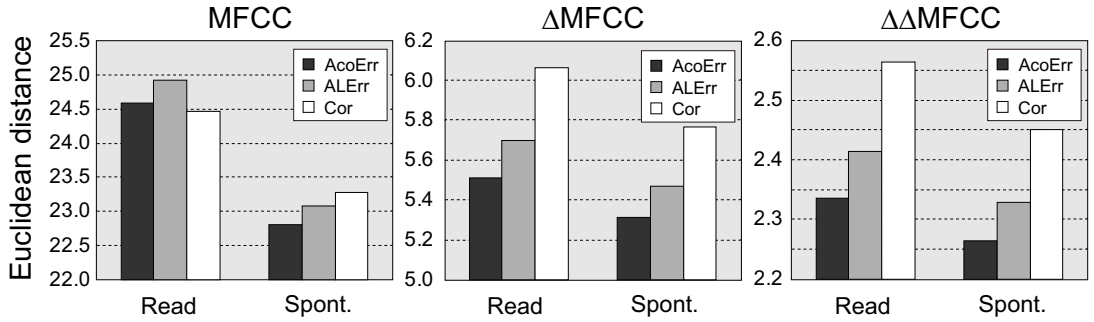

Fig. 2　音響的誤りフレーム，音響・言語的誤りフレーム，正解フレームにおける原点からの距離の違い

Table 1　読み上げ音声と話し言葉音声での各誤り原因の生起頻度と生起率 (%)

	Read		Spont.	
SearchErr	157,268	(19.4)	282,458	(19.6)
ALErr	108,404	(13.4)	199,386	(13.8)
AcoErr	99,729	(12.3)	231,242	(16.1)
LngErr	25,592	(6.1)	52,967	(7.9)
Corr	420,045	(51.8)	674,547	(46.8)

声においてはケプストラム空間の縮小に対する処理を行うことで認識性能向上の可能性があることを示している．

図 2 に，読み上げ音声と話し言葉音声における，音響的誤りフレーム (AcoErr)，音響・言語的誤りフレーム (ALErr)，正解フレーム (Cor) における MFCC, ΔMFCC, ΔΔMFCC の原点からのユークリッド距離の違いを示す．図 2 の左，中央，右の図はそれぞれ，MFCC, ΔMFCC, ΔΔMFCC 12 次元における距離の平均値を示す．各図中の横軸の Read は読み上げ音声，Spont. は話し言葉音声を表す．話し言葉音声の認識精度は 74.9 %，読み上げ音声の認識精度は 79.5 % となった．図 2 より，MFCC, ΔMFCC, ΔΔMFCC 全てにおいて，読み上げ音声に対して話し言葉音声では原点からの距離が小さくなっており，この結果は我々の先行研究と一致する．話し言葉音声の MFCC では読み上げ音声とは異なり，正解フレームと比較して音響的誤りフレームと音響・言語的誤りフレームの原点からの距離が小さくなっていることから，局所的なケプストラム空間の縮小が生じていると考えられる．ΔMFCC，ΔΔMFCC では，読み上げ音声・話し言葉音声の両方で音響的誤りフレームにおける局所的なケプストラム空間の縮小が生じていることが分かる．

5　ケプストラム空間拡張の効果

前節の結果を元に，話し言葉音声における音響的誤りフレームに対してケプストラム空間を拡張させた場合の認識性能の変化に関して分析を行った．これは，音響的誤りフレームが正しく検出できたと仮定した場合にどれだけの認識性能向上が見込めるかを調べることに相当し，局所的なケプストラム空間の縮小が認識性能の劣化にどれだけ影響を及ぼしているかを調べることが出来る．ケプストラム空間の拡張は，MFCC，ΔMFCC，ΔΔMFCC の各要素に対して，それぞれ定数倍することで実現した．図 3 に，音響的誤りフレームにおけるケプストラム空間の拡張処理による認識性能向上への効果を示す．横軸は音響的誤りフレームに対する拡張率 r を表し，1.0 から 2.0 まで 0.05 刻みで適用した．縦軸は単語正解精度を表す．図中の □ は，音響的誤りフレームの MFCC 12 次元のみに対して各拡張率を適用し，それ以外の要素の拡張率を 1.0 とした場合の認識性能の推移を表す．図中の △，◯ はそれぞれ同様に，ΔMFCC 12 次元，ΔΔMFCC 12 次元のみに対して，各拡張率を適用した場合の認識性能の推移を表す．図 3 より，MFCC, ΔMFCC, ΔΔMFCC ではそれぞれ絶対値で 0.29%, 0.59%, 0.33% の向上が見られた．この結果により，認識性能の低下の原因の一つとして，局所的なケプストラム空間の縮小が関係していることが分かり，音響的誤りフレームに対して適切なケプストラム空間の拡張処理を行うことで，認識性能の向上が見込めることが明らかになった．

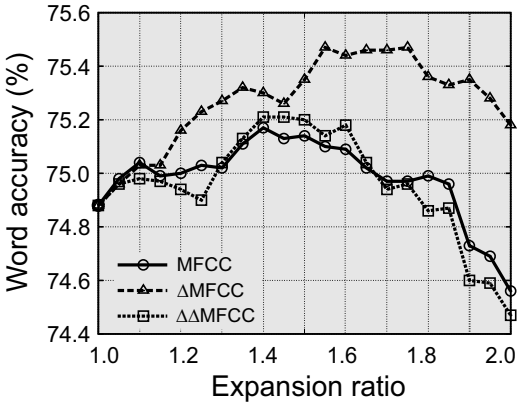

Fig. 3　音響的誤りフレームにおけるケプストラム空間の拡張処理による認識性能向上への効果

6　おわりに

本研究では，各フレームに対して認識誤り原因の同定を行うことで音響的誤りフレームを検出し，ケプストラム空間の縮小について分析を行った．その結果，音響的誤りフレームでは，正解フレームに対して原点からの距離が小さくなっていることが明らかになった．さらに，これらのフレームに対して MFCC, ΔMFCC，ΔΔMFCC の各要素に対してケプストラム空間の拡張処理を行ったところ，認識性能の向上が見られた．これにより，認識性能の低下の原因の一つとして局所的なケプストラム空間の縮小が関係していることが分かった．すなわちこの結果は，音響的誤りフレームに対してケプストラム空間の適切な拡張処理を行うことで，認識性能が向上する見込みがあることを意味している．今後の課題としては，局所的なケプストラム空間の縮小への対策の一つとして，音響的誤りフレームにおいてケプストラム空間を拡張させるための，より適切な方法を考案する必要がある．また，本分析では正解単語列を用いることで音響的誤りフレームの検出を行っているため，誤り同定を行わずに音響的誤りフレームを正確に検出する方法を考案する必要がある．

参考文献

[1] M. Nakamura, et al., “Analysis of spectral space reduction in spontaneous speech and its effects on speech recognition performances,” *Proc. Interspeech 2005*, pp.3381–3384, 2005.

[2] M. Nakamura, et al., “The effect of spectral space reduction in spontaneous speech on recognition performances,” *Proc. ICASSP2007*, SPE-L7.5, 2007.

[3] 南條ら,「大語彙連続音声認識における認識誤り原因の自動同定」, 情報処理学会研究報告, 99-SLP-27-6, pp.41–48, 1999.