論文 / 著書情報
Article / Book Information

論題(和文)	KLDを用いた中国語における読み上げ音声と話し言葉音声の違いの分析
Title(English)	
著者(和文)	中村 匡伸, 劉 鵬, 宋 謌平, 古井 貞熙
Authors(English)	Masanobu Nakamura, SADAOKI FURUI
出典(和文)	日本音響学会 2007年秋季講演論文集, Vol. , No. 2-4-1, pp. 323-324
Citation(English)	, Vol. , No. 2-4-1, pp. 323-324
発行日 / Pub. date	2007, 9

KLDを用いた中国語における読み上げ音声と話し言葉音声の違いの分析*

◎中村 匡伸[1], 劉 鵬[2], 宋 謌平[2], 古井 貞熙[1]
(1: 東工大, 2: Microsoft Research Asia)

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) に収録されている同一話者の発声した話し言葉音声にお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]．

これまで我々は日本語の音素を分析の対象にしてきたが，本稿では中国語の音素を対象とし，読み上げ音声と話し言葉音声の違いに関して分析を行う．これにより，日本語に限らない話し言葉音声の一般的な特徴を明確化できる可能性がある．また本分析では Kulback-Leibler 擬距離を用いて 2 音素間の音響モデルの距離を算出し，認識性能との関係を明らかにする．

2 音声データ

分析には，話し言葉音声としては電話での対話音声を用い，読み上げ音声としては，同一話者の発声したニュース記事読み上げ音声，および朗読音声を用いた．これらは全て中国語で発声され，pinyin 表記の書き起こしがなされている．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，無音区間で区切られた約 10 秒程度の区間を「発話単位」として定義する．

3 音響モデルの学習

本実験において分析対象とする音素は，四声および軽声を区別した全 184 種類の中国語音素とした．本分析を行うにあたり，読み上げ音声，話し言葉音声における混合ガウス monophone HMM を作成する．

1. 音声データから MFCC 12 次元と対数パワー，その一次微分および二次微分成分の計 39 次元の音響パラメータを抽出する．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．
2. 読み上げ音声，話し言葉音声の各発話スタイルごとに，学習用データを用いて混合ガウス monophone HMM を学習する．全ての音素モデルは，3 状態の left-to-right 型 HMM とする．

表 1 に，音響モデルの学習および評価に用いた発話単位数と発話時間を示す．学習には 460 名 (男女各 230 名) の話者，評価には 30 名 (男性 16 名，女性 14 名) の話者を用いた．

4 Kulback-Leibler 擬距離

Kulback-Leibler 擬距離 (以下 KLD とする) は，2 つの確率密度関数の類似度を算出する擬距離である．我々は以前にマハラノビス距離を用いて 2 つの単一ガウス分布関数の間の距離を求めた．KLD は，一般的な 2 つの確率密度関数の距離を算出することが可能な，汎用性の高い距離尺度である．本分析では，2 つの HMM の間の KLD を近似的に算出する．

Table 1 音声データの発話単位数と発話時間

	発話スタイル	発話単位数	発話時間 (時間)
学習	話し言葉	38,298	148.4
	読み上げ	27,586	66.9
評価	話し言葉	2,610	9.8
	読み上げ	1,799	4.4

KLD の定義式は通常 closed-form による解を持たないため，モンテカルロ法を用いて近似計算を行う必要がある．2 つの GMM $s, \tilde{s}$ の KLD は unscented transform[2] を用いて以下のような近似式で表される [3].

$$D(s \parallel \tilde{s}) \approx \frac{1}{2N} \sum_{m=1}^{M} \omega_m \sum_{k=1}^{2N} \log \frac{p(\boldsymbol{o}_{m,k} \mid s)}{p(\boldsymbol{o}_{m,k} \mid \tilde{s})} \qquad (1)$$

ただし，N は音響特徴量の次元数 ($N = 39$), M は混合数，ω_m は GMM における m 番目の Gaussian kernel の混合重み，$\boldsymbol{o}_{m,k}(1 \leq k \leq 2N)$ は m 番目の Gaussian kernel の k 番目の sigma point とする．sigma point には，$\boldsymbol{o}_{m,k} = \boldsymbol{\mu}_m + \sqrt{N\lambda_{m,k}}\boldsymbol{u}_{m,k}$, $\boldsymbol{o}_{m,k+N} = \boldsymbol{\mu}_m - \sqrt{N\lambda_{m,k}}\boldsymbol{u}_{m,k}$ $(1 \leq k \leq N)$ を選んだ．ここで $\boldsymbol{\mu}_m$ は m 番目の Gaussian kernel の平均ベクトル，$\lambda_{m,k}, \boldsymbol{u}_{m,k}$ は m 番目の Gaussian kernel の共分散対角行列における k 番目の固有値と固有ベクトルである．

本分析では，全ての GMM の間の KLD を算出し，動的計画法 (Dynamic Programming) および状態遷移確率を用いて 2 つの monophone HMM の間の KLD を近似的に求めた [4].

5 全音素間の KLD の分布

本分析では，読み上げ音声と話し言葉音声において音響モデルの混合数を増やした場合の全音素間の KLD の変化を比較する．音響モデルの学習には，表 1 に示す学習用話者を用いて，発話スタイルごとに混合数 1, 2, 4, 8, 16, 32, 64 の monophone HMM を作成する．

学習済みの monophone HMM において 2 つの音素 $p_i, p_j (i \neq j)$ の間の KLD を求め，音素 p_i における k 近傍の値を用いて相対累積度数分布を作成する．図 1 に，読み上げ音声と話し言葉音声の全音素間の KLD の分布を示す．図中の分布は全て，各音素で 10 近傍法を用いて作成した．左図に読み上げ音声における全音素間の KLD の分布，右図に話し言葉音声における全音素間の KLD の分布を示す．x 軸は KLD を表し，y 軸は相対累積度数を表す．読み上げ音声における各混合数の KLD の分布の中位値は，それぞれ 11.8, 15.0, 17.4, 18.8, 20.8, 22.8, 26.0，話し言葉音声における各混合数の KLD の分布の中位値は，それぞれ 11.1, 12.9, 14.5, 15.8, 16.9, 18.0, 19.9 となった．これより，音響モデルの混合数が増えるにしたがって全音素間の KLD が大きくなることが明らかになった．さらに，読み上げ音声に比べて話し言葉音声では全音素間の KLD が小さく，混合数を増や

*Analysis of the difference between read and spontaneous Mandarin Chinese with KLD, by Masanobu Nakmaura[1], Peng Liu[2], Frank K. Soong[2], and Sadaoki Furui[1] (1: Tokyo Institute of Technology, 2: Microsoft Research Asia)

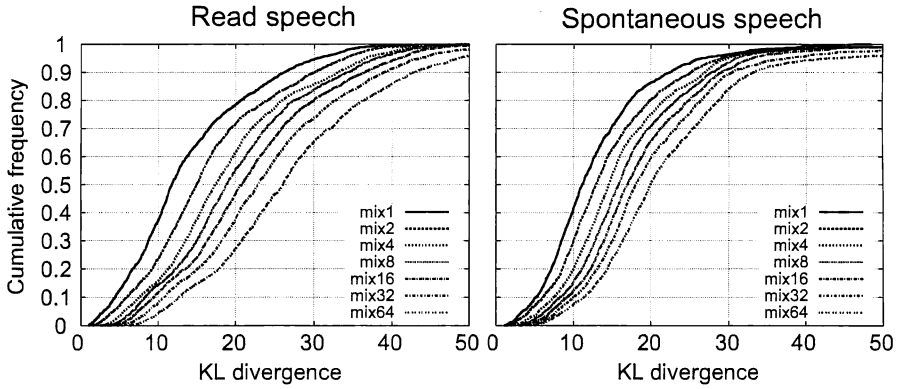

Fig. 1　読み上げ音声と話し言葉音声の全音素間の KLD の分布

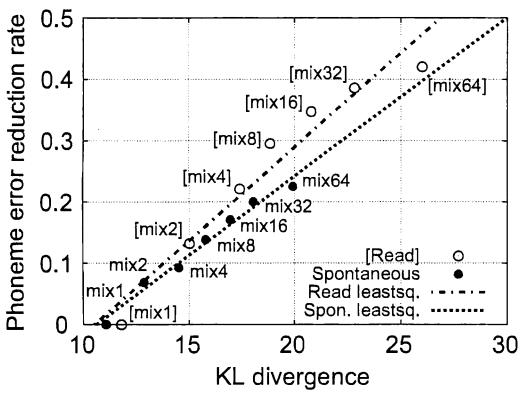

Fig. 2　KLD と音素誤り改善率との関係

しても全音素間の KLD の増加率が小さいことが明らかになった．

6　全音素間の KLD と音素認識精度との関係

我々の先行研究によれば，全音素間のマハラノビス距離の平均値と音素認識性能の間には強い相関関係があることが明らかになっている [1]．そのため，本分析においても全音素間の KLD と音素認識性能との関係を導く．

音響モデルとしては前節で分析を行った 1, 2, 4, 8, 16, 32, 64 混合 monophone HMM を用いる．評価対象の音声データは，表 1 に示す 30 名の音声データを用いる．これらは学習データには含まれていない．本実験では音節ベースのネットワークを用いて音素認識を行い，音素正解精度を算出する．音素認識を行う際には，挿入ペナルティは発話スタイルごとに最適な値を用いる．本分析では音素正解精度の代わりとして音素誤り改善率 (Per) を用いる．音素誤り改善率は，混合数 m の音素正解精度 $Acc(m)$ を用いて以下の式により算出される．

$$Per(m) = \frac{\big(100 - Acc(1)\big) - \big(100 - Acc(m)\big)}{100 - Acc(1)} \quad (2)$$

図 2 に，KLD と音素誤り改善率との関係を示す．x 軸は各発話スタイルの混合数ごとの音響モデルにおける全音素間の KLD の中位値を表し，y 軸は音素誤り改善率を表す．図中の○は読み上げ音声，●は話し言葉音声を表す．図中の読み上げ音声の各混合数の表記は [ ] で囲み，話し言葉音声の表記と区別した．読み上げ音声と話し言葉音声における KLD と音素誤り改善率との相関係数はそれぞれ 0.97, 0.99 となった．図中の直線は，最小二乗法を用いて各発話スタイルにおける点列を一次関数で近似したものである．

これらの結果により，全音素間の KLD と音素認識性能との間には強い相関関係があることが明らかになった．さらに，読み上げ音声に対して話し言葉音声では，全音素間の KLD が小さいため音素認識性能が低く，混合数を増やしても音素誤り改善率の増加があまり見られない，という特徴が見られることから，話し言葉音声における認識性能の改善の難しさが定量的に示された．

7　まとめ

本分析では KLD を用いて中国語における読み上げ音声と話し言葉音声の違いに関する分析を行った．結果として，音響モデルの混合数を増やすことで全音素間の KLD が増加することが明らかになり，読み上げ音声に対して話し言葉音声では全音素間の KLD の増加の割合が小さくなっていることが明らかになった．さらに全音素間の KLD の大きさと音素認識性能との関係を調べたところ，0.97 以上の強い相関関係があることが明らかになった．また読み上げ音声に対して話し言葉音声では，全音素間の KLD が小さいため音素認識性能が低く，混合数を増やした時の音素誤り改善率が小さいことが明らかになった．このことは，話し言葉音声の認識性能の改善の難しさを定量的に示している．

本分析では monophone モデルを用いて全音素間の KLD の違いを明らかにしたが，通常の音声認識で用いられる tri-phone モデルにおいても同様の分析を行う必要がある．本分析では中国語の音素を分析対象としたが，日本語の音素に関しても同様の分析を行いたい．また，今回得られた知見を，話し言葉音声の認識性能の向上に役立てることができるかどうか，検討する必要がある．

謝辞　本研究は文部科学省 21 世紀 COE プログラム「大規模知識資源の体系化と活用基盤構築」の一環として実施されました．

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] J. Goldberger, et al., "An efficient image similarity measure based on approximations of KL-divergence between two gaussian mixtures," *Proc. International Conference on Computer Vision 2003*, pp. 370–377, 2003.

[3] J. Du, et al., "Minimum divergence based discriminative training," *Proc. Interspeech 2006*, pp. 2410–2413, 2006.

[4] P. Liu, et al., "Divergence-based similarity measure for spoken document retrieval" *Proc. ICASSP 2007*, pp. IV-89–IV-92, 2007.