社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2004－SLP－53　(2)
2004／10／22

日本語話し言葉コーパスを用いた話し言葉音声の音響的特徴の分析

中村匡伸　　岩野公司　　古井貞熙

東京工業大学大学院 情報理工学研究科 計算工学専攻
〒 152-8552 東京都目黒区大岡山 2-12-1
Email: {masa, iwano, furui}@furui.cs.titech.ac.jp

本稿では、日本語話し言葉コーパスに収録されている話し言葉音声と読み上げ音声の音響的特徴を比較する。話し言葉音声として学会講演音声、模擬講演音声、対話音声を取り上げ、読み上げ音声として学会講演音声の書き起こしを読み上げた再読み上げ音声を取り上げる。これらの音声データは、共通の話者が異なる発話スタイルで発声しているため、話者によるスペクトルの違いの影響が除去できる。男女各 5 名の音声データを用いて音素ごとにケプストラムの分布の縮小率を求め、比較を行ったところ、話し言葉音声のケプストラム空間は、読み上げ音声に対して縮小するという特徴を持っており、中でも対話音声はその特徴が顕著に現れることが分かった。男女各 3 名の音声データを用いて発話速度の分布を比較したところ、話し言葉音声の発話速度の分布は、再読み上げ音声に対して発話速度が大きくなることが分かった。

Analysis of acoustic characteristics in sponaneous speech using Corpus of Spontaneous Japanese

Masanobu Nakamura, Koji Iwano, and Sadaoki Furui

Department of Computer Science, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
Email: {masa, iwano, furui}@furui.cs.titech.ac.jp

This paper compares acoustic characteristics of spontaneous speech and read speech analyzed using the Corpus of Spontaneous Japanese (CSJ). Academic presentations, extemporaneous presentations and dialogue utterances were analyzed as spontaneous speech, and the utterances reading manual transcriptions of academic presentations were analyzed as read speech. Since they were uttered in different speaking styles by a group of common speakers, the problem of individual spectral difference could be avoided. Reduction of cepstral distribution for each phone was analyzed using utterances by 5 male and 5 female speakers. It was found that the cepstral distribution of spontaneous speech was reduced comparing to that of read speech. This was especially significant for dialogue utterances. Speaking rate was also analyzed using phonetically labeled utterances spoken by 3 male and 3 female speakers, and it was found that the speaking rate of spontaneous speech was higher than that of read speech.

1　はじめに

近年、書き言葉の読み上げ音声の認識においては高い認識精度が得られるようになってきている。しかし一方で、話し言葉音声などいわゆる自然発話の音声認識を行うと、その認識精度は極端に低下する。話し言葉音声の分析を行うことで、読み上げ音声とは異なる話し言葉特有の音響的な性質を明確化することができれば、話し言葉音声の認識性能の向上に役立つ知見が得られると期待される。そこで本稿では、日本語話し言葉コーパス (以降では CSJ と呼ぶ) に収録されている話し言葉音声と読み上げ音声の音響的な特性の比較を行う。これらの音声データは、共通の話者が話し言葉音声・読み上げ音声の双方を発声しており、話者によるスペクトルの違いの影響を除去することが可能である。

音響的特徴としては、各音素のケプストラムと発話速度に注目する。我々はすでに [1]、CSJ に収録されている学会講演音声と、同一話者による同一内容の読み上げ音声を用い、各音素のケプスト

ラムの比較を行った。その結果、読み上げ音声に対する話し言葉音声のケプストラム空間が縮小するという傾向が得られた。また同様のデータを用いて発話速度の分布の比較を行ったところ、読み上げ音声に対する話し言葉音声の発話速度が大きいという結果が得られた。

本稿では、分析対象の話し言葉に模擬講演音声、対話音声を加え、ケプストラムと発話速度の観点から同様の傾向が得られるか調査する。

2　音声データ

本実験で用いた音声データはCSJに収録されている男性・女性話者各5名による読み上げ音声と話し言葉音声である。話し言葉音声として学会講演音声、模擬講演音声、対話音声を用い、読み上げ音声として学会講演音声の書き起こしを読み上げた再読み上げ音声を用いる。対話音声としては学会講演のインタビュー、模擬講演のインタビュー、自由対話、課題対話を用いる。

表1に本実験で用いた音声データの音素サンプル数を示す。再読み上げ音声、学会講演音声、模擬講演音声、対話音声をそれぞれR,A,S,Dとしている。10名の話者の中にはCSJのコアに含まれる話者が男女3名ずつおり、これらの音声データには、人手によって音素ごとに時間ラベルが付加されている。

音声データは16kHzでサンプリングされた、それぞれ約15分のデータである。実験に際して、まず転記ファイルを元に講演音声を400ms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。

3　ケプストラムに関する分析

3.1　ケプストラムの分布の縮小率

読み上げ音声と話し言葉音声の比較は、読み上げ音声に対する話し言葉音声のケプストラムの分布の縮小率で行う。発話タイプ X の音素 p におけるケプストラムの分布の縮小率 $red_p(X)$ を次のように定義する。

$$red_p(X) = \frac{\|\boldsymbol{c}_p(X) - \overline{\boldsymbol{c}_p(X)}\|}{\|\boldsymbol{c}_p(R) - \overline{\boldsymbol{c}_p(R)}\|}$$

再読み上げ音声、学会講演音声、模擬講演音声、対話音声の発話タイプをそれぞれ R, A, S, D とし、

表2: 音素のリスト

母音	/a,i,u,e,o,a:,i:,u:,e:,o:/
子音	/w,y,r,p,t,k,b,d,g,j,ts,ch, z,s,sh,h,f,N,N:,m,n/

$\boldsymbol{c}_p(X)$ を発話タイプ X の音素 p における平均ケプストラムとする。また発話タイプ X の全音素の平均ケプストラムを $\overline{\boldsymbol{c}_p(X)}$ とする。話者ごとに $red_p(X)$ を求め、その話者平均値を $\overline{red_p(X)}$ とする。ノルムにはユークリッド距離を用いる。

3.2　平均ケプストラムの抽出

ケプストラムの分布の縮小率の算出には、話者・発話タイプごとに求めた各音素の平均ケプストラムの値を用いる。対象とした音素は、表2のリストにある31種（母音10種・子音21種）とした。また、分析対象データは表1にあげた全音声データ（40種類）とした。

ここで各音素の平均ケプストラムは、以下のようにして抽出される。

(1) 音声データからMFCC12次元と対数パワー、それらの一次微分と二次微分成分の計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) 各話者・発話タイプごとに、分析対象データを用いて1混合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。

(3) 出来上がったmonophone HMMのうち、分析対象音素のHMMの第2状態から12次元MFCCのベクトルを音素の平均ケプストラムとして取り出す。

4　発話速度に関する分析

話し言葉音声の発話速度は読み上げ音声と比較して変動幅が大きく、その影響で認識誤りが生じやすいという調査結果がある[2]。そのため、話し言葉音声と読み上げ音声の発話速度分布の傾向を調査することは、認識性能の向上に有効であると考える。

発話タイプによる発話速度の違いを調べる実験

表 1: 音声データの音素サンプル数

	話者 ID	コア	再読み上げ音声 (R)	学会講演音声 (A)	模擬講演音声 (S)	対話音声 (D)
男性	M1	○	7,420	7,371	5,213	9,915
	M2	○	10,768	10,815	6,000	14,489
	M3	○	12,118	12,211	8,525	17,616
	M4	–	23,154	23,208	8,615	19,892
	M5	–	8,598	8,651	11,518	29,862
女性	F1	○	12,162	12,071	10,119	25,428
	F2	○	7,843	7,757	7,206	20,141
	F3	○	11,383	11,360	4,837	17,044
	F4	–	8,111	8,038	8,232	20,999
	F5	–	17,797	17,848	9,598	22,083

には、表 1 のコアに含まれる 6 名の音声データを用いた。付与されている音素ラベルを用いて、各音素の継続時間長を求めた後、その逆数を取ることで各音素の発話速度を算出する。単位は音素/secとなる。分析対象の音素は、3.1 節と同じく、表 2 に示す音素とした。

5 実験結果

5.1 ケプストラムの分布の縮小率

図 1 に、母音と子音のケプストラムの分布の縮小率 $\overline{red_p(X)}$ を発話タイプごとに示す。上図が母音、下図が子音を表している。$\overline{red_p(X)} = 1$ を太線で表記する。

図 1 の左側に再読み上げ音声に対する学会講演音声のケプストラムの分布の縮小率 $red_p(A)$ を表す。これより、母音における縮小率はほとんど 1 に近いことが分かる。これは、母音のケプストラム空間の広がりを見たとき、再読み上げ音声と学会講演音声の間に差がないことを示している。逆に子音における縮小率は、大半の音素が 1 よりも小さくなっている。ただし、音素/ch/,/N:/に関しては、用いた音声データでは発声されなかったため縮小率は求めていない。図 1 の中央に再読み上げ音声に対する模擬講演音声のケプストラムの分布の縮小率 $red_p(S)$ を表す。学会講演音声と比較すると、母音における縮小率が若干小さくなっている。図 1 の右側に再読み上げ音声に対する対話音声のケプストラムの分布の縮小率 $red_p(D)$ を表す。学会講演音声と比較すると、母音においても子音においても縮小率が小さくなっている。

また図 2 に母音と子音の $\overline{red_p(X)}$ の平均を発話タイプごとに示す。学会講演音声、模擬講演音声、対話音声をそれぞれ A, S, D とした。これより、全ての発話タイプでケプストラムの分布の縮小が生じており、対話音声ではそれが顕著にあらわれている。

ケプストラム空間の縮小の様子を視覚的に表す目的で、ある話者における対話音声と再読み上げ音声の 12 次元の MFCC を 2 次元の主成分ベクトル空間に射影した結果を図 3 に示す。各点は各音素に対応し、上図が母音、下図が子音の分布を表している。第 1、第 2 主成分ベクトルをそれぞれ横軸、縦軸に取っている。再読み上げ音声の各音素の平均ケプストラムを白点で表し、その分布を破線の楕円で近似している。また対話音声の各音素の平均ケプストラムを黒点で表し、その分布を実線の楕円で近似している。これより、全ての音素の平均ケプストラムが分布の中心 (0,0) に近づく、すなわち音素によるケプストラムの違いが小さくなっている様子が分かる。

また、以下のような補足実験を行った。音声データとしてコアに含まれる男女各 3 名の音声を用い、付与されている音素ラベルを用いて、各音素区間を切り出し、各音素の 12 次元 MFCC の平均値を求める。この平均ケプストラムを用いて、再読み上げ音声に対する各発話タイプのケプストラムの分布の縮小率を求める。結果を図 4 に示す。上図は母音、下図は子音の分析結果である。HMM の第 2 状態から平均ケプストラムを抽出した場合の結果（図 1）と同様に、3 種の話し言葉音声についてケプストラム空間が縮小し、特に対話音声における

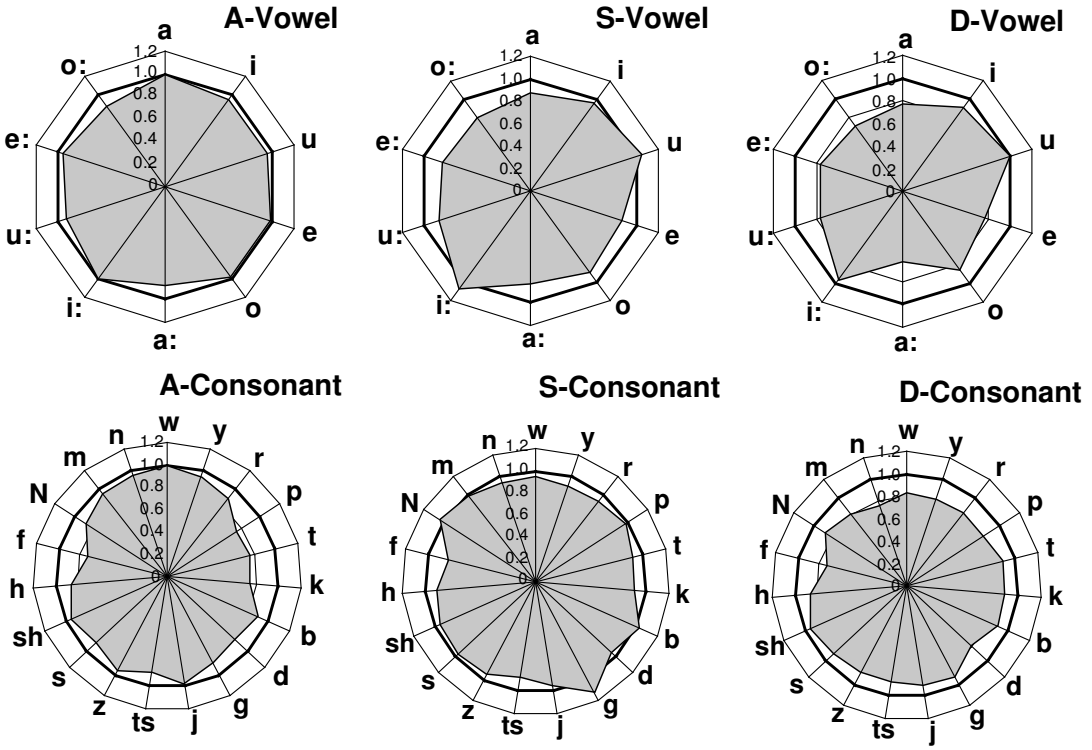

図 1: 各発話タイプにおける母音・子音のケプストラムの分布の縮小率 (左から学会講演音声 (A)、模擬講演音声 (S)、対話音声 (D))

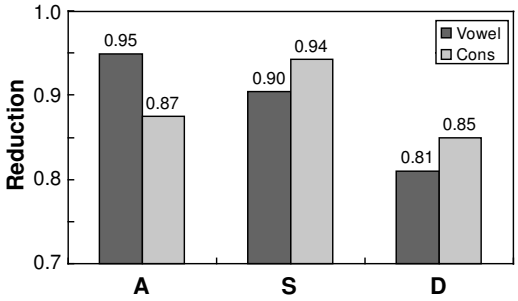

図 2: ケプストラムの分布の縮小率の平均

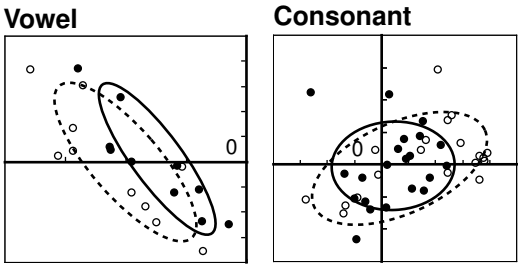

図 3: 対話音声における母音・子音のケプストラムの分布

空間の縮小度合いが大きい、という傾向が見られる。補足実験における平均ケプストラムの抽出法は、人手によって付与された音素の時間ラベルを利用していることから、正確な分析結果を得ることができるが、ラベル付きデータが大量に存在しない場合には、結果がばらつく危険性が伴う。HMM を用いる平均ケプストラムの抽出方法は、データ量の確保が容易であり、かつ、人手によるラベルを用いた補足実験と同様の傾向が得られたことから、本分析において妥当な特徴抽出手法であったということができる。

5.2 発話速度

図 5 に、各発話タイプの発話速度分布を示す。再読み上げ音声の発話速度分布とその中位値を破線で表し、学会講演音声、模擬講演音声、対話音声の発話速度分布とその中位値を実線で表す。上図は

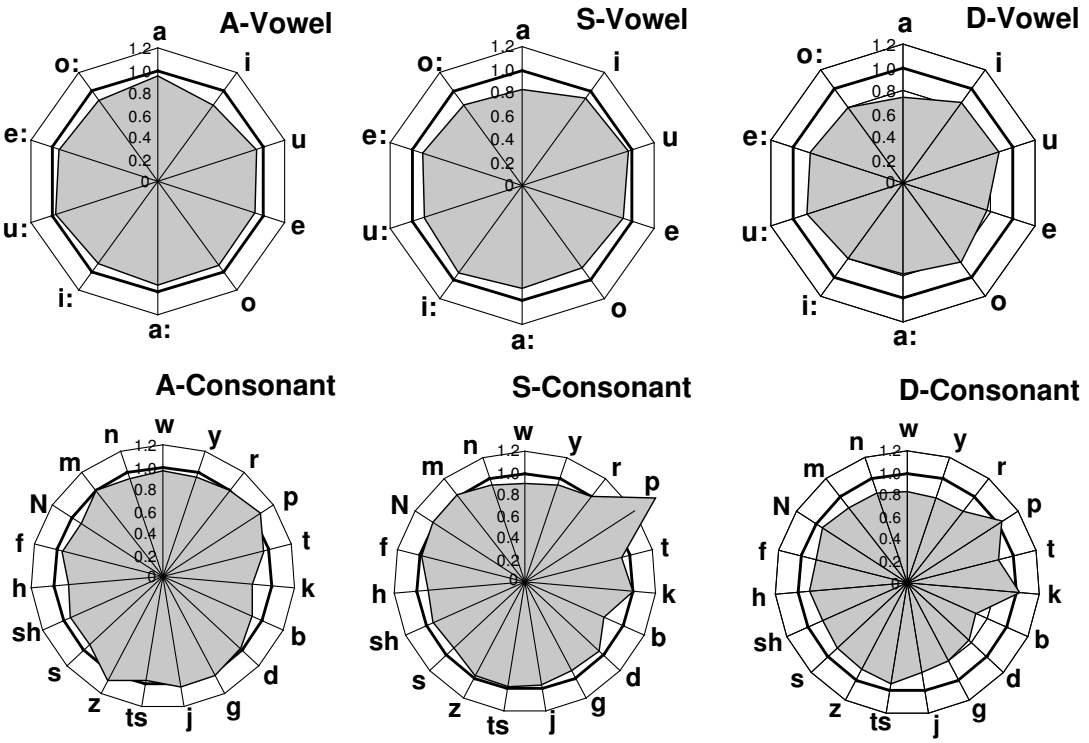

図 4: 音素ラベルを用いて各音素区間を切り出した時の各発話タイプにおける母音・子音のケプストラムの分布の縮小率 (左から学会講演音声 (A)、模擬講演音声 (S)、対話音声 (D))

母音、下図は子音の分布を表す。また図 6 に発話タイプによる母音・子音の発話速度分布の中位値を示す。再読み上げ音声、学会講演音声、模擬講演音声、対話音声をそれぞれ R, A, S, D とした。これより、再読み上げ音声に対して、学会講演音声、模擬講演音声、対話音声の発話速度が大きくなっていることが分かる。また、読み上げ音声と話し言葉音声の発話速度分布の中位値に有意差があるかどうかを検定するため、ウィルコクソンの順位和検定 (U 検定) を行った。ここでは計算量の関係で、各発話タイプについて無作為に 1,500 音素を抽出し、分布を再構成した後に検定を行っている。その結果、再読み上げ音声と対話音声の子音以外については読み上げ音声に対して話し言葉音声の発話速度が大きいと言うことが有意水準 1%で確認された。

6 まとめ

本実験では、話し言葉音声と読み上げ音声の音響的な特性の違いを明らかにするため、CSJ のデータを用いて読み上げ音声と発話スタイルの異なる話し言葉を、各音素のケプストラムと発話速度の 2 つに注目して分析した。

再読み上げ音声に対し、学会講演音声、模擬講演音声、対話音声のケプストラム空間の縮小度合いはどれも大きくなる特徴を持ち、特に対話音声ではその特徴が顕著に現れることが分かった。

発話速度に関しては、母音、子音ともに読み上げ音声よりも話し言葉音声の方が大きくなる傾向が確認された。この結果は、音声データとして CSJ の学会講演音声と ATR の読み上げ音声を比較した研究報告 [3] とも一致している.

本稿で取り上げたケプストラムの分布の縮小率・発話速度は、音声の「自発性」を示す指標であると考えられる。この仮定に一般性があるかどうか調べるためには、CSJ 以外の他のコーパスに対しても同様の分析を行う必要がある。ただし、本実験では「同一話者による話し言葉音声と読み上げ音声」を対象データとしているため、そのようなデータが存在しない、他のコーパスを分析する前に、話者の違うデータを扱ったときの縮小率の変動について CSJ を用いて調査しておく必要がある。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出来るかどうか、検討する必要もある。

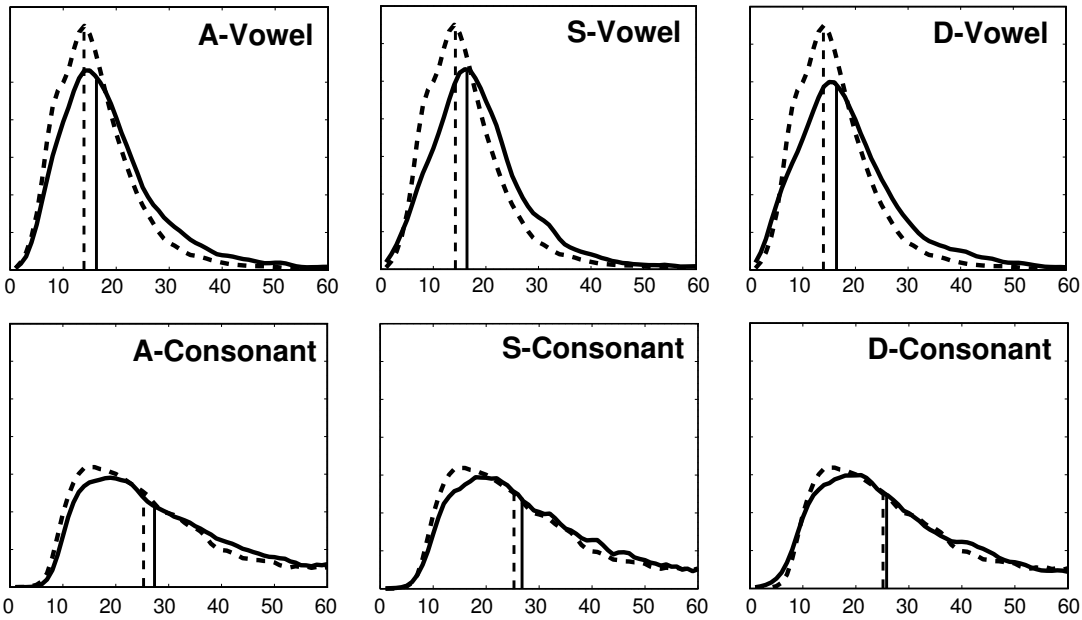

図 5: 各発話タイプの発話速度分布 (左から学会講演音声 (A)、模擬講演音声 (S)、対話音声 (D))

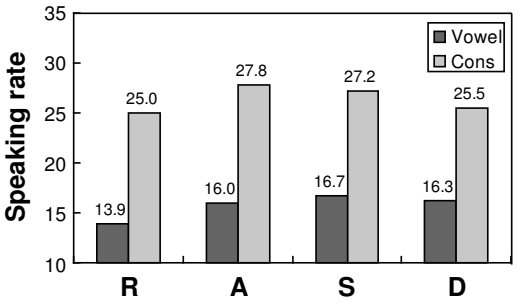

図 6: 発話速度分布の中位値

謝辞

本研究を進めるに当たり，貴重なご助言をいただいた前川喜久雄氏 (国語研) に感謝いたします．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] 中村匡伸, 岩野公司, 古井貞熙, “日本語話し言葉音声と読み上げ音声の音響的特徴の比較,” 日本音響学会秋期研究発表会講演論文集, 2-P-5 (2004-9).

[2] 篠崎隆宏, 古井貞熙, “話し言葉認識における決定木を用いた誤り要因の分析,” 日本音響学会研究発表会講演論文集, 1-1-9 (2001-10).

[3] K.Maekawa, “Corpus of Spontaneous Japanese: Its Design and Evaluation,” *Proc. SSPR 2003*, pp.7-12 (2003-4).