# 電話音声の話者認識における雑音とハンドセットの影響*

6 N − 4

宮本宗易　滝口哲也　中村　哲　鹿野清宏

（奈良先端科学技術大学院大学 情報科学研究科）

## 1　はじめに

電話音声の話者認識を行なう時、認識性能を劣化させる要因がいくつかある。一つは電話機固有の周波数特性や電話回線の回線特性等（以後、チャネル特性と表現）による音声の歪みである。もう一つは、外部から入る対象話者以外の音声や電話を使用している際の外部環境の雑音（以後、外部雑音と表現）による音声の劣化である。特に、電話音声で話者認識をする場合、異なる電話機を利用することが考えられ、その場合でも頑健に認識を行なう必要がある。しかし、電話機や電話回線（以後、ハンドセットと表現）が異なるとチャネル特性が変わるため頑健な認識ができない。そこで、本稿では異なるハンドセットから入力された電話音声にいろいろな手法を適用し、認識実験によりそれらの効果を調べる。

## 2　外部雑音の除去

電話音声に重なる外部雑音により、話者認識性能は劣化してしまう。これら外部雑音を効果的に除去する方法としてSS(Spectral Subtraction)[1]がよく使われる。SSとは、パワースペクトル領域において雑音区間で逐次推定した雑音のパワースペクトルを入力スペクトルから減算を行なうことである。つまり、

$$\hat{N}(\omega;t)=\begin{cases}\hat{N}(\omega;t-1) & \text{音声区間}\\ \gamma\cdot\hat{N}(\omega;t-1)+(1-\gamma)\cdot O(\omega;t) & \\ & \text{それ以外}\end{cases} \quad (1)$$

$$\hat{S}(\omega;t)=\max(O(\omega;t)-\alpha\cdot\hat{N}(\omega;t),\beta\cdot O(\omega;t)) \quad (2)$$

である。ここで、$\alpha,\beta,\gamma$はSSパラメータと呼ばれる実定数である。$\hat{N}(\omega;t),\hat{S}(\omega;t),O(\omega;t)$は、それぞれ時刻$t$の推定後の雑音、推定後の信号、観測信号であり、いずれもパワースペクトルである。

## 3　チャネル特性の正規化

同一文章を発話してもハンドセットが異なると電話音声の特徴ベクトルは全く異なったものとなる。そこで、ハンドセットの違いを取り除くためケプストラム平均正規化法（CMN:Cepstral Mean Normalization）[2]によりチャネル特性の正規化を試みる。この方法は、観測信号のケプストラムベクトルの平均を各観測ベクトルから減算する方法である。つまり、

$$\hat{O}_{mean}=\frac{1}{T}\sum_{t=1}^{T}O_{cep}(t) \quad (3)$$

$$\hat{O}_{CMN}(t)=O_{cep}(t)-\hat{O}_{mean} \quad (4)$$

である。ここで、$O_{cep}(t)$および$\hat{O}_{CMN}(t)$はそれぞれ時刻$t$の観測信号およびCMN後の観測信号のケプストラムベクトル、$\hat{O}_{mean}$は観測信号のケプストラムの平均である。なお、$T$は観測信号長である。ここでは、更に次の式に示すように観測信号$O_{cep}(t)$の標準偏差$\sigma$でベクトルの各要素を正規化（Norm）する[3]。

$$\sigma_i=\sqrt{\frac{1}{T}\sum_{t=1}^{T}(O_i-\mu_i)^2} \quad (5)$$

$$\hat{O}_i=\frac{O_i-\mu_i}{\sigma_i} \quad (6)$$

ここで、$\mu_i,\sigma_i,O_i,\hat{O}_i$はそれぞれ特徴ベクトルの$i$番目要素の平均、標準偏差、観測信号、正規化後の観測信号である。

## 4　話者認識実験

実験条件を表1に示す。データベースはSPIDREを使用した[4]。これは英語発話の電話音声を収録した話者認識のためのデータベースである。この中から対象話者45人（男27、女18）がさまざまなハンドセットのなかから3種類を使って会話したものを使用した。話者モデルを作成するため各話者から適当に一つ選択した会話の音声部分の初めから60秒を使った。また、評価データとしてモデル作成のデータと同じハンドセットで会話内容の異なるデータ(Task1)と、モデル作成で使わなかった残りの2会話(Task2)の音声部分の初めから30秒を使用した。まず、ハンドセットが異なることにより認識性能がどれだけ劣化するか調べるため評価データにTask1とTask2を使った。その実験結果が表2である。この結果から異なるハンドセットの場合、認識性能が

*"Effects by noise and handsets in speaker recognition of telephone speech", by M. Miyamoto, T. Takiguchi, S. Nakamura, and K. Shikano (Graduate School of Information Science, Nara Institute of Science and Technology)

表 1: 話者認識実験条件

| | |
|---|---|
| データベース | SPIDRE |
| サンプリング | 8bit $\mu$ low |
| フレーム幅／シフト | 32msec ／ 8msec |
| 特徴パラメータ | MelCep16 次<br>+ΔMelCep16 次 |
| 話者数 | 45 人 |
| 話者モデル次数 | 1 状態 64 混合ガウス分布 |

半分以下にまで劣化し、話者認識が困難になることが示された。これはハンドセットが異なったことにより違うチャネル特性が評価データに影響し、話者モデルの分布とのずれが性能劣化を導いてしまったと考えられる。

表 2: ハンドセットの影響 [%]

| タスク | 認識率 |
|---|---|
| Task1（同一ハンドセット） | 77.8 |
| Task2（異なるハンドセット） | 34.4 |

次に、データに CMN および SS を適用し SS パラメータの$\alpha$を変化させ、Task2 を評価データとして使って認識性能の変化を調べた。なお、$\beta,\gamma$はそれぞれ 0.1,0.974 に固定して実験を行なった。その時の認識結果を図 1 に示す。同図より$\alpha$が 2.8 のとき最も良い認識率を示した。$\alpha < 2.8$ の時は音声区間で十分に雑音が除去されなかったと考えられる。一方、$\alpha > 2.8$ の時は前のとは逆に音声区間で除去した雑音が大き過ぎたと考えられる。このことからこれ以降の実験では、$\alpha = 2.8$ として実験を行なっている。

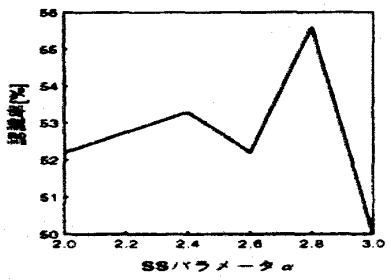


図 1: CMN および SS を適用した認識率と$\alpha$の値

さらに、CMN,SS,Norm を適用することによりどれだけの改善が得られるか調べる。評価データに Task2 のみを使用し、SS パラメータは $\alpha$=2.8,$\beta$=0.1,$\gamma$=0.974 としている。データには、雑音を除去するために SS を、チャネル特性を正規化するために CMN を、さらに標準偏差によるデータのばらつきの正規化をそれぞれ施し、それらの効果を文献 [5] で行なわれた実験と比較する。この結果を表 3 に示す。まず、SS を適用したことにより同表中の 1,2 段目や 3,4 段目の比較から 4%強の改善が見られた。これは、わずかであるが SS によって音声部分の雑音の除去ができていることによると考えられる。また、同表中の 1,3 段目や 2,4 段目を比較すると CMN によるチャネル特性の正規化の効果で 17%の改善が得られていることが分かる。また、MIT[5] で行なわれた実験結果の約 55%と CMN+SS の認識結果がほぼ同等であった。最後に、同表中の 5,6 段目に Norm を施した時の結果を示す。この結果より Norm を施すことによりさらに認識性能を改善することができた。特に、CMN+Norm のときもっとも良い認識率を得た。

表 3: 各処理に対する話者認識実験結果 [%]

| 処理 | 認識率 |
|---|---|
| なし | 34.4 |
| SS | 38.9 |
| CMN | 51.1 |
| CMN+SS | 55.6 |
| CMN+Norm | 60.0 |
| CMN+SS+Norm | 58.9 |

## 5　おわりに

本稿では、電話音声が持つ問題点、外部雑音の存在とチャネル特性による歪みについて述べた。そして、それらを補償する手段として SS,CMN および Norm を適用し、それらが有効な手段であることを実験により示した。

## 参考文献

[1] Boll. S., "Supression of acoustic noise in speech using spectral subtraction", IEEE Trans, ASSP-27,no.2,pp.113–120,1979

[2] B.Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic spaker identification and verification", J. Acoust. Soc. Amer., Vol.55, pp 1304–1312, 1974

[3] Olli ら, "Noise Robust HMM-based Speech Recognition Using Segmental Cepstral Feature Vector Normalization", April, ESCA-NATO Workshop, pp 107–110, 1997

[4] John J. Godfrey ら, "SWITCHBOARD: Telephone speech corpus for research and development", ICASSP pp. I-517-I-520, 1992

[5] Douglas A. Reynolds, "The Effects of Handset Variability on Speaker Recognition Performance:Experiments on THe Switchboard Corpus", ICASSP pp. I-113-I-116, 1996