講義音声認識精度改善のためのチョーク音除去

張　安 † 西崎博光 ‡ 関口芳廣 ‡
山梨大学大学院 †医学工学総合教育部・‡医学工学総合研究部

1 はじめに

筆者らは，講義音声の認識の研究を行っている[1]．これまでの経験から，黒板を使った講義音声の認識精度は，プロジェクタ等を使用した講義の認識精度より低いことがわかっている．その原因の１つはチョーク音の影響である[2]．

本研究では，講義音声から突発的な不定常雑音である独立したチョーク音を除去して，講義音声認識精度の向上を目指す．

2 チョーク音除去の方法

チョーク音は周期性を持たず，チョークの使い方により音の長さや周波数成分が一定しないため，特徴を掴むことが難しい．そこで，音声データから人間の声を検出し，人間の声以外の音を取り除くことによってチョーク音を除去する方法を提案する．

人間の声を検出するために，大きく分けて下記の２つのパラメータを使用する．

1. 周波数成分の特徴(PA)：音の周波数スペクトル（0～8 kHz）
2. 波形の特徴（PB)：音の波形を使用する（単位時間当たりの振幅の変化量を使用）

そして，PA と PB の線形結合によって，音声らしさ Y を求め，Y によってチョーク音か否かを判定する．

2.1 周波数成分の特徴（PA)

周波数成分の特徴について説明する．

2.1.1 基本周波数とその倍音の明確さ（A_1)

図１は，音声の周波数スペクトル（0～2.6 kHz）の概形である．基本周波数とその倍音の明確さを式(1)で求める．ここで，n は(0～2.6kHz)帯のスペクトルピークの数，

$$A_1 = \sum_{n=1}^{n} \left(V_{(i)} - P_{(i)}\right)^2 \qquad (1)$$

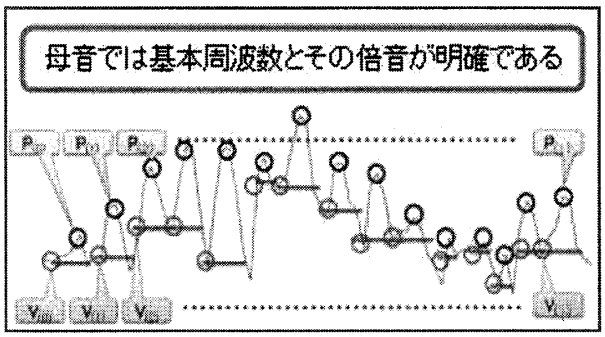

図１　周波数スペクトルの概形（一部分）

$V_{(i)}$ と $P_{(i)}$は，基本周波数とその倍音のピークの始点と頂点である．音声(母音)は，基本周波数とその倍音がはっきりしており，チョーク音は，それがはっきりしていない．そのため，音声の場合は，A_1が大きくなり，チョーク音や無音の場合は，A_1が小さくなる．A_1は，音の大きさと高さにより若干の変動はあるが，講義のようなきちんとした発音であれば，経験的に一般にA_1の値は大きくなる．

2.1.2 低周波成分と高周波成分の比率（A_2)

一般に音声の低周波成分は高周波成分より大きい．チョーク音は低周波成分と高周波成分の差が小さい．この特徴を利用することで，音声とチョーク音を判別する．

高周波成分と低周波成分の比 A_2 は式(2)のように計算できる．A_L と A_H はそれぞれ式(3)，式(4)で求める．A_Lは，0～2.6kHz 帯のパワースペクトルのうちで最大のものをとり，A_Hには 5.3～8.0kHz 帯の中で最大のパワースペクトル値が設定される．A_2 が大きいと音声，小さいとチョーク音である可能性が高い．

$$A_2 = \frac{A_L}{A_H} \qquad (2)$$

$$A_L = \max\left\{SP_{(k)} \middle| 0.0 \le k \le 2.6\text{kHz}\right\} \qquad (3)$$

$$A_H = \max\left\{SP_{(k)} \middle| 5.3 \le k \le 8.0\text{kHz}\right\} \qquad (4)$$

2.1.3 周波数成分の特徴の計算（PA)

最終的に，周波数成分の特徴 PA は式(5)で計算する．

$$PA = A_1 \times A_2 \qquad (5)$$

基本周波数とその倍音の明確さを，低・高周波成分の比で強調させる．PA は，音声部分では大きな値をとることになる．

2.2 波形の特徴（PB)

波形の特徴を PB で表す．PB は下記の B_1 と B_2に分けて考える．B_1 は波形の変化の激しさ，つまり単位時間当たりの振幅の変化量である．B_2 は音の長さ，つまり音の継続時間長に関係する特徴である．

2.2.1 波形の変化の激しさ（B_1)

図２の例のように，チョーク音のような突発的な不定常雑音は，瞬間的なパワーが大きいので単位時間あたりの振幅の変化量が大きい．隣同士のデータの差分をとったものが図３となる．図３に示すように，チョーク音は音声と比べて，差分値が大きくなる．図３に対して音声を 10m 秒の矩形窓で切り出しフレーム内の最大値を $B_{0(i)}$とする．ただし，i = 1～m で，m は継続したフレームの数

* Chalk noise removal for improvement of lecture speech recognition.
by An Zgang, Hiromitsu Nishizaki and Yosihiro Sekiguti(University of Yamanashi)

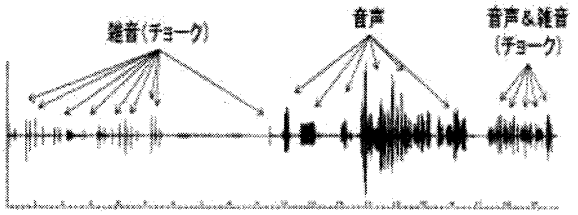

図 2　講義音声データの波形の一部分(例)

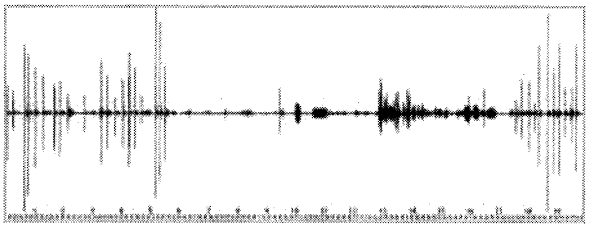

図 3　図2の講義音声データの波形の差分(例)

である.

2.2.2　音の継続時間 (B_2)

これまでの調査で，チョーク音の発生継続時間はほぼ125m秒以下であることが分かっている．連続した125m秒以上のものをまとめ，音の継続時間(B_2)とする．10m秒単位毎に音の継続時間(B_2)を求める.

2.2.3　波形の特徴の計算 (PB)

式(6)である音の継続時間に対応する単位時間当たりの差分波形の和の平均値 B_1 を求める．波形の特徴は式(7)に示すように，B_1 と B_2 の比で定義する．チョーク音では単位時間あたりの振幅の変化量が大きく，継続時間が短いので，PB の値が大きくなる．音声の場合は逆なので，PB の値が小さい.

$$B_1 = \frac{1}{B_2}\sum_{i=1}^{m} B_{0(i)} \qquad (6)$$

$$PB = \frac{B_1}{B_2} \qquad (7)$$

2.3　音声らしさの値 (Y)

式(8)に示すように，PA と PB それぞれに α と β の重みをつけて線形結合を行い，音声らしさの値 Y を求める.

$$Y = (PA \times \alpha) + (PB \times \beta) \qquad (8)$$

Y の値は，音声の方が大きく，チョーク音の方が小さくなる．しきい値を設定し，Y の値に基づいて音声とチョーク音を判別する.

3　チョーク音除去実験

チョーク音除去の効果を確かめるための実験を行った．チョーク音を含む講義音声から提案手法によりチョーク音の除去を行い，その結果を音声認識することで評価を行う．使用した音声認識エンジンなどの実験条件は次の通りである.

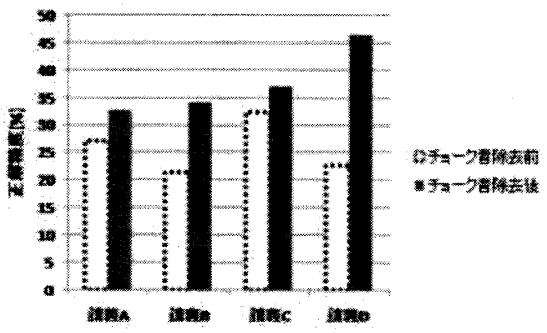

図 4　チョーク音除去前後の音声認識正解精度の比較

・処理対象音声：
　山梨大学工学部で収録した4講義(A,B,C,D).
・音声認識エンジン：Julius4.1
・言語モデル・辞書：
　講義音声の書き起こしから学習したトライグラムモデル.
　(認識対象の講義は含まない)
・評価：単語正解精度で評価.
・Y算出用重み：α=2, β=-1
・Yのしきい値：チョーク音< 60 <音声
　(なお，α，βとYのしきい値は実験的に求めた)

図 4 に実験結果を示す．チョーク音を除去する前は，チョーク音が誤認識され何らかの単語に認識され挿入誤りとなっていた．除去することにより，大幅に挿入誤りが減少し，その影響で正解率も改善できている．しかし，チョーク音が単体で出現している部分は良く除去できているが，音声とチョーク音が混在している部分の除去が難しいことが明らかとなった．また，子音部分が誤って除去される現像も確認できた.

4　まとめ

本稿では，講義音声に含まれるチョーク音除去手法の提案を行った．チョーク音の周波数成分の特徴と波形(音の変化量)の特徴を利用することで，チョーク音を削除できることが分かった.

提案手法による講義音声中のチョーク音削除により，講義音声認識性能を改善することに成功した．しかし，音声とチョークの混在部分に関しては，音声情報までも削除してしまうこともあり，今後の改善が必要である．また，講義中に含まれる突発的な他の雑音についても検討していきたい.

参考文献

[1] 藤原裕幸，西崎博光，関口芳廣，“話題依存言語モデル構築のための LSA と単語発話情報を用いた語彙推定”，第 8 回情報科学技術フォーラム講演論文集，第 2 分冊，RE-004，pp.35--42，2009.9
[2] 小林健司，西崎博光，関口芳廣，”講義音声を対象とした音声評価と認識率の関係”，日本音響学会，2008年秋季研究発表会講演論文集，2-P-18，pp.397-298，2008.9