クリッピングした音声の音声認識*

◎太刀岡勇気, 成田知宏, 石井純 (三菱電機・情報総研)

1　はじめに

音声認識の実用化に伴い、録音条件の悪い環境で収録された音声を認識しなければならない場面が増加している。音声を劣化させる要因として、不適切なゲインの設定に起因するクリッピングの問題がある。クリッピングにより明らかに音声は劣化するが、それがどの程度音声認識性能に悪影響を与えるかは明らかにされていない。そこで本報では様々なクリッピングレベルのクリッピング信号を人工的に作成し、音声認識率との関係を評価した。また Hermite 補間による簡易的な波形復元法を検討した。

2　クリッピング信号と原信号の推定方式

クリッピング信号とは振幅を-1 から 1 で正規化した信号 y に対して、Eq. (1) で表されるクリッピングレベル θ_c でクリッピングさせた信号 y_c のことである。

$$y_c = \begin{cases} \mathrm{sign}(y)\theta_c & (|y| \geq \theta_c), \\ y & (|y| < \theta_c). \end{cases}$$

ここで sign は引数が正の場合に 1 を、負の場合に-1 を返す関数である。

2.1　既存の推定方式

クリッピング信号の回復法はいくつか提案されている。例えば文献 [1] の自己回帰による未観測信号の推定手法を、クリッピング信号の推定問題に適用した場合に良好な SNR 改善が得られることが示されている [2]。また、予め定義した基底を選択する方法も提案されており、例えば [2] ではスパース性を仮定して Orthogonal Matching Pursuit アルゴリズムにより近似解を求める手法が、[3] では逐次ベクトル射影法を用いた手法が示されている。

2.2　Hermite 補間による簡易な推定方式

上記の原信号の推定手法は計算負荷が大きい。クリッピングの頻度は実際にはそれほど高くないので、起こったときと起こらないときの計算負荷の差が小さい方が望ましく、音声認識の前処理として用いるためには簡便な手法が求められる。ここでは Hermite 補間により y_c から $\hat{y}$ を内挿する方法について述べる。

ここでは正の値でクリップが起こった場合を考察するが、負の場合も全く同様である。クリッピング

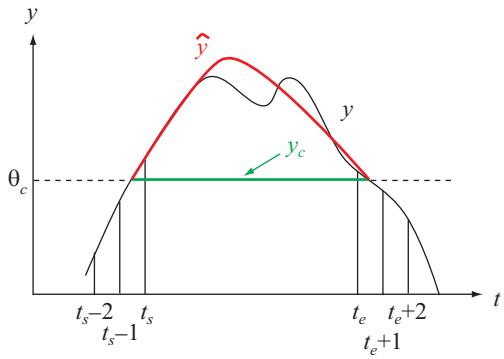

Fig. 1　Hermite interpolation of clipped waveform.

区間の端点を Fig. 1 のように t_s,t_e とし、$y(t_s-1)$ と $y(t_e+1)$ を通る 3 次関数を内挿する。その際に微分値 y' の算出が問題となる。ここでは

$$\begin{aligned} y'(t_s-1) &\approx \underline{y(t_s-1)-y(t_s-2)} \\ &\approx y(t_s)-y(t_s-1) \geq \underline{y_c(t_s)-y(t_s-1)} \end{aligned} \quad (1)$$

$$\begin{aligned} y'(t_e+1) &\approx \underline{y(t_e+2)-y(t_e+1)} \\ &\approx y(t_e+1)-y(t_e) \leq \underline{y(t_e+1)-y_c(t_e)} \end{aligned} \quad (2)$$

の関係を利用して、Eq. (1) の右辺下線部の大きい方と Eq. (2) の右辺下線部の小さい方を微分値とすることで、3 次関数を内挿する。

3　音声認識実験

3.1　実験条件

評価データは、電子協 100 地名のデータベース (16kHz サンプリング) を Eq. (1) の θ_c を 0.2 から 0.9 まで変化させて作成した。話者は男女 20 名ずつとした。

評価尺度としては Eq. (3) で表される SNR と PESQ[4]、音声認識率 (単語) で評価した。

$$SNR = 10\log\frac{\sum y^2(t_c)}{\sum\left(y(t_c)-\hat{y}(t_c)\right)^2} \quad (3)$$

t_c はクリッピングしているサンプルの時間である。PESQ は音声品質を評価する客観指標 (評価値は 0.5 ～4.5) で [4]、雑音環境下において音声認識率との対応がよいことが示されている [5]。音声認識率の評価では、タスクの難易度によるクリッピングの影響を考察するために、タスク難易度を認識語彙を変えること

*Speech recognition of clipped speech, by TACHIOKA, Yuuki, NARITA, Tomohiro, ISHII, Jun (Mitsubishi Electric Corp.).

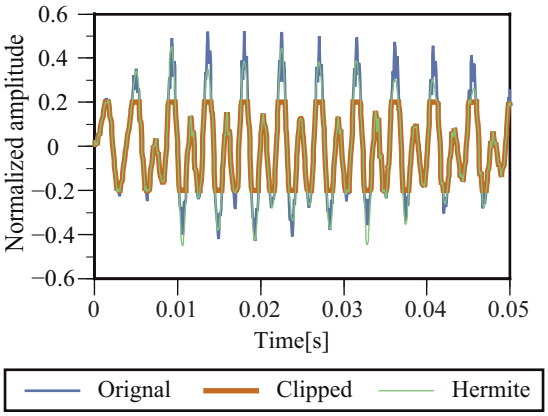

Fig. 2　Original, clipped, and restored waveforms.

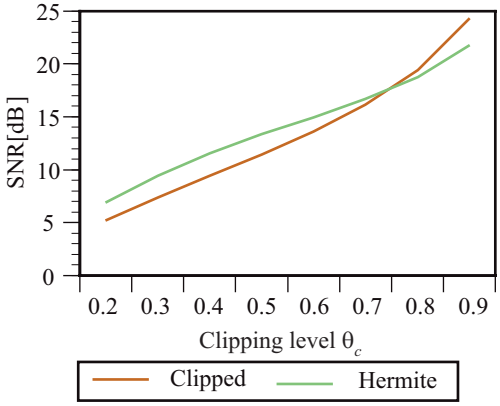

Fig. 3　SNR before and after restoration.

で難 (155,592 単語) と易 (100 単語) の 2 パターン設定した。音響モデルは、julius に付属の 3131 状態、64 混合 PTM(8256 ガウス分布) の tri-phone HMM を用い、julius(ver.4.2.1) により認識実験を行った。音響特徴量は、12 次元の MFCC とその Δ に Δ パワーを加えた 25 次元の特徴量とした。

3.2　結果と考察

まず Hermite 補間による波形の復元程度を Fig. 2 に示す。Hermite 補間によってある程度元の波形を予測できていることがわかる。SNR の改善量を Fig. 3 に示す。クリッピングレベルが低い場合に約 3dB の SNR の改善がみられる。

PESQ と認識率の結果を Fig. 4 に示す。クリッピングレベル θ_c と PESQ は順序通りに並んでいる。タスクにも依存するが θ_c が 0.7 程度までは認識率の低下は少なく、聴感上もそれほど音質劣化を感じなかった。それより θ_c が小さくなると認識率は低下し、タスクが難しいものの方が認識率は著しく低下した。点線は、Yamada[5] の提案するロジスティック分布の最小 2 乗法によるデータフィッテングであり、PESQ と認識率の関係をよく説明できている。

波形復元による認識率の改善を Fig. 5 に示す。θ_c が 0.2 から 0.5 の平均で認識率が 2.5%改善している。タスク難の場合も同傾向であった。

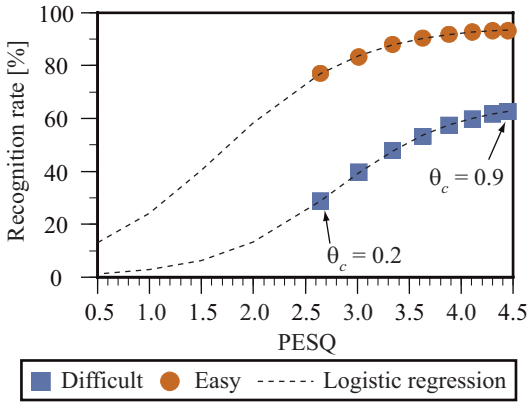

Fig. 4　Recognition rate (easy task (dictionary: 100 words) and difficult task (dictionary: 155,592 words)) and PESQ.

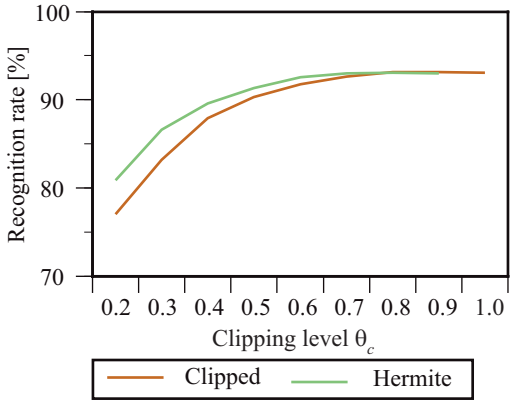

Fig. 5　Recognition rate before and after restoration (Easy task).

4　まとめと今後の課題

クリッピングした音声の認識実験を行った。Hermite 補間に基づく簡易的な波形復元法により、クリッピングレベルが低い場合に SNR と認識率の向上が見られた。また雑音下音声認識において認識率と相関が高い PESQ がクリッピング音声の認識においても認識率と相関が高い可能性が示された。

今後の課題としては、クリッピングに強い音声特徴量や、音声認識に適したクリッピング信号の回復法、大語彙連続音声認識での PESQ によるクリッピング音声の認識率予測の検討があげられる。

参考文献

[1] A. Janssen *et al.*,*IEEE Trans. on ASSP*, **34**, 317–330 (1986. 4).
[2] A. Adler *et al.*,*ICASSP 2011*, 329–332 (2011).
[3] S. Miura *et al.*,*TENCON 2011*, 787–790 (2011).
[4] ITU-T Rec. P862, *http://www.itu.int/rec/T-REC-P.862/*, (2001).
[5] T. Yamada *et al.*, *IEEE Trans. on ASLP*, **14**, 2006–2013 (2006. 11).