音声認識のための LSTM-RNN を用いた帯域拡張

BAND-WIDTH EXTENSION USING LSTM-RNN FOR SPEECH RECOGNITION

太刀岡勇気　Yuuki Tachioka　　石井純　Jun Ishii

三菱電機 (株) 情報技術総合研究所　Information Technology R&D Center, Mitsubishi Electric Corporation

1　はじめに

広帯域音声を使うことで音声認識の性能は向上するが，音響モデル学習時と異なる狭帯域の音声が入力されると音声認識性能は著しく低下する．主に聴感上の主観評価を向上させることを目的に，帯域拡張技術が検討されている．モデルベースの手法としては，ガウス混合分布 (Gaussian Mixture Model; GMM) によるものがある [1]．近年，再帰的な構造を持つ神経回路網 (Recurrent Neural Network; RNN) の性能を向上させた Long-short term memory(LSTM)-RNN [2] が，auto encoder として高い再構築性能を持つことが知られている．本報では帯域拡張に LSTM-RNN を用い，その性能を音声認識により評価した．

2　GMM に基づく帯域拡張法

GMM に基づく声質変換手法 [1] を応用して，帯域拡張を行う．帯域拡張前の音声を元話者，拡張後の音声を対象話者と考え，帯域拡張前後の特徴量の結合ベクトルを全共分散 GMM でモデル化する．特徴量はメルケプストラム (mcep) と MFCC の 2 通り検討した．mcep は，8kHz の音声から 17 次元，16kHz の音声から 25 次元を抽出し，その Δ 特徴量を加えた計 84 次元の結合ベクトルを用いた．mcep から波形を復元し，MFCC を抽出して音声認識を行った．MFCC の場合には，8kHz,16kHz いずれの場合にも 13 次元とし，その Δ 特徴量を加えた計 52 次元の結合ベクトルを用いた．この場合は得られた MFCC を直接用いて，音声認識を行った．実験には SPTK toolkit(ver.3.7) を利用した．

3　LSTM-RNN に基づく帯域拡張法

mcep の場合，8kHz 音声の 17 次元+Δ(34 次元) から，16kHz 音声の 25 次元+Δ(50 次元) を予測する LSTM-RNN を，2 乗誤差最小基準で学習した．MFCC は，13 次元+Δ の 26 次元のベクトルを入出力とした．実験には currennt toolkit (ver.0.2) を利用した．

4　音声認識実験による各手法の評価

Kaldi toolkit を利用し，TIMIT の音素認識で評価した．GMM の音響モデルで認識し，特徴量は MFCC+Δ+Δ^2 とした．加えて LDA+MLLT による特徴量変換と fMLLR による話者適応を行った．表 1 に開発セットでのベースラインを示す．16kHz の音声を 16kHz のモデルで認識した場合が認識性能が高く，特に話者適応を行った場合に差が広がっている．次によいのが 8kHz の音声を 8kHz のモデルで認識した場合である．モデルと音声のサンプリング周波数が異なる場合は，ある程度適応で補えるものの，認識性能が著しく低下している．

表 1　Phoneme error rate (PER) [%] on the **dev** set, evaluating 16kHz and 8kHz speech. 8kHz speech were recognized by using 16kHz and 8kHz models.

	MFCC	+LDA+MLLT	+fMLLR
16k (16k model)	23.1	21.3	18.9
8k (16k model)	32.3	28.8	23.4
8k (8k model)	23.5	22.5	20.3

表 2　PER [%] on the **dev** set, evaluating GMM and LSTM-RNN based band-width expansion.

	MFCC		+LDA+MLLT		+fMLLR	
Feature	mcep	mfcc	mcep	mfcc	mcep	mfcc
GMM	29.3	30.4	27.9	29.0	25.4	24.7
LSTM	25.5	24.7	23.9	23.0	22.1	20.7

表 3　PER [%] on the **test** set.

	MFCC		+LDA+MLLT		+fMLLR	
16k	24.9		22.3		19.9	
8k	34.8		30.4		25.3	
8k (Matched)	25.1		23.5		21.0	
Feature	mcep	mfcc	mcep	mfcc	mcep	mfcc
GMM	31.4	32.6	29.8	29.9	26.6	26.1
LSTM	27.2	25.9	25.2	24.0	23.4	21.9

次に帯域拡張を行った場合の結果を，表 2 に示す．予備実験で，声質変換の従来研究を参考に，性別依存と性別非依存 2 種類のモデルを試したが，それらの差異は小さかったため，性別非依存のメルケプストラム/MFCC 両方の場合の結果を載せている．すべての場合で，GMM を LSTM-RNN が上回っている．MFCC を直接抽出した場合には，GMM の話者適応を用いない場合に性能が低下したが，それ以外の場合は性能が向上した．音声認識が目的の場合は，音声認識に適した特徴量を直接推定することが有効であることがわかる．帯域拡張により，モデルを切り替えることなく，8kHz の音声の認識性能を向上させることができた．表 3 に，テストセットでの評価結果を示す．傾向は開発セットでの場合と同様で，LSTM-RNN による帯域拡張は有効であった．

5　おわりに

LSTM-RNN による帯域拡張は GMM よりも効果が大きく，MFCC を直接予測する方法が有効であった．

参考文献

[1] T. Toda et al., IEEE Trans. on Audio, Speech, Language Process, **15**, 2222–2235, 2007.

[2] S. Hochreiter and J. Schmidhuber, Neural Computation, **9**, 1735–1780, 1997.