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

2005－ＳＬＰ－59（33）
2005／12／22

音声特徴抽出法 SPADE を用いたフロントエンドの耐雑音評価標準コーパスによる評価

石塚 健太郎　中谷 智広

NTT コミュニケーション科学基礎研究所，日本電信電話株式会社
〒619-0237 けいはんな学研都市 精華町 光台 2-4
E-mail: {ishizuka, nak}@cslab.kecl.ntt.co.jp

あらまし　音声特徴抽出法 SPADE と耐雑音信号処理・特徴量正規化処理を併用した，耐雑音フロントエンドを提案し，その性能を雑音下音声認識評価環境 AURORA-2J で評価した．耐雑音信号処理として非線形スペクトルサブトラクションや適応ウィーナフィルタを用い，特徴量正規化処理としてケプストラム平均正規化や分散正規化を用いた．結果，Clean Training 条件で最大 82.58 %, Multicondition Training 条件で最大 92.55 %の単語正解精度を得た．

キーワード　音声特徴量，耐雑音フロントエンド，帯域分割，周期性，非周期性

Evaluation of Front-End Based on Speech Feature Extraction Method SPADE Using Standard Noise Robustness Evaluation Corpus

Kentaro ISHIZUKA and Tomohiro NAKATANI

NTT Communication Science Laboratories, NTT Corporation
Hikaridai 2-4, Seikacho, Keihanna Science City, 619-0237, Japan
E-mail: {ishizuka, nak}@cslab.kecl.ntt.co.jp

Abstract　This paper proposes noise-robust front-end processing combining speech feature extraction method 'SPADE' with noise suppression and feature normalization methods. A nonlinear spectral subtraction or an adaptive Wiener filter method was adopted as noise suppression. The cepstral mean or variance normalization method was adopted as feature normalization methods. As results from the evaluation using AURORA-2J, the proposed method achieves word accuracies of 82.58 % for clean training condition, and 92.55 % for multicondition training condition, at highest.

Keyword　Speech feature, Noise robust frontend, Subband, Periodicity, Aperiodicity

1. はじめに

MFCC 等の従来の声特徴は音響変動に弱く，より頑健な音声特徴表現が必要である．我々はこれまで，入力信号を帯域分割し，各帯域内で周期性成分と非周期性成分を分離し，音声特徴表現として併用する音声特徴抽出法 SPADE (Sub-band based Periodicity and Aperiodicity DEcomposition)を提案し[1, 2]，雑音下音声認識評価環境 AURORA-2J[3]での評価により頑健性を示した．また，SPADE の処理を全て周波数領域で実現し[4]，スペクトル減算法[5]や RASTA[6]などの，時間-周波数領域で実現される耐雑音信号処理技術との併用が容易となった[7]．

本稿では，音声特徴抽出法 SPADE と，周波数領域処理の雑音抑圧手法とを併用した耐雑音フロントエンドについて述べ，AURORA-2J を用いた評価結果を報告する．

2. SPADE を用いた耐雑音フロントエンド

本稿で提案する耐雑音フロントエンドは，音声特徴抽出法 SPADE と，特徴抽出前に行う耐雑音信号処理，並びに特徴抽出後に行う特徴量正規化処理を併用し構成される．また，対数パワーパラメータの代わりに用いる周期性成分対数パワーを提案する．以下，各処理について述べる．

2.1. 音声特徴抽出法 SPADE

Fig. 1 に周波数領域での SPADE の処理を示す．本手法では，まず入力信号を Hanning 窓で分析，離散フーリエ変換（DFT）しスペクトログラムを求める．次に，スペクトログラムに帯域通過フィルタバンクの周波数特性を乗じ，複数の周波数帯域に分割する．その後，各帯域のパワースペクトルを離散逆フーリエ変換（IDFT）し自己相関関数を求め，信号の基本周期を推定する．この推定基本周期に基づき櫛型フィルタを帯域ごとに独立に設計し，その周波数特性を帯域分割されたスペクトルに乗じてフィルタリングを行う．この櫛型フィルタによって主要な周期成分が抑圧されたスペクトルの対数パワーを合算したものを非周期性成分とみなし，一方で櫛型フィルタにより抑圧された分の対数パワーを周期性成分とみなして，周期性成分と非周期性成分を分離する．最後に，同時刻フレームの周期成分と非周期成分を帯域横断的に連結してベクトル化し，それらを離散コサイン変換して得られる係数をさらに連結し一つのベクトルとする．このベクトルを音声認識の特徴パラメータとして用いる．

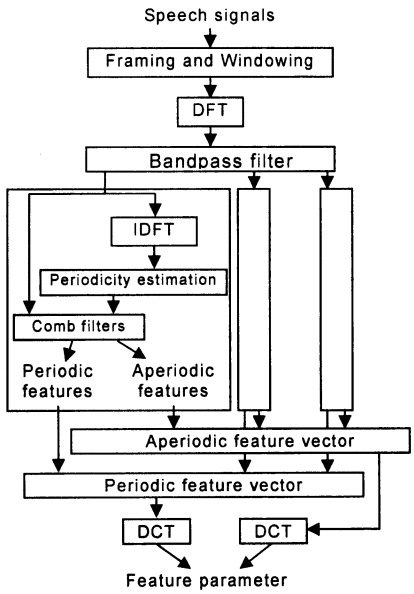

Fig. 1: Block diagram of SPADE.

2.2. 耐雑音信号処理

スペクトログラムに適用する耐雑音信号処理として，Berouti らの非線形スペクトル減算(NSS)[8]，または Adami らの適応ウィーナフィルタ(AWF)[9]を用いた．これらは加法性歪を軽減する．また，周期性・非周期性成分の各ベクトルに対し，RASTA[6]を適用した．RASTA は音声固有の振幅特徴のみを抽出し，乗法性歪の軽減効果もある．

2.3. 特徴量正規化処理

離散コサイン変換後得られる SPADE の特徴量の正規化処理として，ケプストラム平均正規化[10]，および分散正規化[11]を行った．いずれも乗法性歪の軽減効果がある．

2.4. 周期性成分対数パワー

SPADE の周期性成分は，音声の振幅変動情報に寄与する有声音部分の特徴を強く反映し，また雑音による変動が小さいと考えられる．この特性を利用するため，音声認識時の特徴パラメータとして通常用いられる信号の対数パワーパラメータの代わりに，周期性成分のパワーを合算しその対数を取った周期性成分対数パワーを用いた．

3. 実験

提案法の効果を，AURORA-2J[3]を用い評価した．評価カテゴリは 3 である．帯域通過フィルタバンクにフィルタ数 24 のガンマトーンフィルタバンクを用い，フレーム長 25 ms，フレームシフト 10 ms，離散コサイン変換後の係数は周期性・非周期性各成分 12 次元の計 24 次元とし，周期性成分対数パワーパラメータを加え，動的特徴としてΔ，$\Delta\Delta$を加えた計 75 次元を特徴ベクトルとした．タスクは連続発声 10 数字認識，識別器は 16 状態 24 混合の数字 HMM であった．

Table 1: Experimental results evaluating robustness of the frontend processing with AURORA-2J. These are the average word accuracies between SNRs of 0 to 20 dB for each test set and all test sets (Overall).

Clean Training (% Accuracy)				
	Test Set			Overall
	A	B	C	
AURORA-2J Baseline	46.51	43.98	49.90	46.17
SPADE	60.45	54.06	65.56	58.92
ETSI WI008 Frontend	79.20	77.81	75.87	77.98
SPADE+NSS+CMN	78.42	78.63	77.73	78.37
SPADE+AWF+CMN	**82.54**	**82.86**	82.12	**82.58**
SPADE+AWF+CVN	81.97	81.21	82.17	81.71
SPADE+AWF+RASTA	82.46	80.22	**82.99**	81.67
Multicondition Training (% Accuracy)				
	Test Set			Overall
	A	B	C	
AURORA-2J Baseline	91.53	80.39	85.83	85.93
SPADE	90.72	83.59	86.52	87.03
ETSI WI008 Frontend	93.23	88.98	90.63	91.01
SPADE+NSS+CMN	93.40	90.00	92.48	91.86
SPADE+AWF+CMN	93.46	**90.81**	92.81	92.27
SPADE+AWF+CVN	**94.48**	90.48	**92.86**	**92.55**
SPADE+AWF+RASTA	93.38	90.71	92.43	92.12

Baseline 評価結果，および ESTI WI008 Advanced DSR Frontend[12]による評価結果[3]と共に，主な結果を Table. 1 に示す．提案法の結果は全ての場合で ETSI WI008 を上回り，全平均認識精度において，Clean training 条件で最大 20.9 %，Multicondition training 条件で最大 17.2 %の単語誤り率削減を達成した．この結果から，SPADE を用いた耐雑音フロントエンドの有効性が示された．

文　献

[1] K. Ishizuka and N. Miyazaki, Proc. ICASSP, vol. 1, pp. 141-144, 2004.

[2] K. Ishizuka, N. Miyazaki, T. Nakatani, and Y. Minami, Proc. Interspeech, vol. 2, pp. 937-940, 2004.

[3] S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa, N. Kitaoka, T. Nishiura, A. Sasou, M. Mizumachi, C. Miyajima, M. Fujimoto, and T. Endo, IEICE Trans. on Inf. & Syst., vol. E88-D, pp. 535-544, 2005.

[4] K. Ishizuka and T. Nakatani, Proc. HSCMA2005, pp. a13-a14, 2005.

[5] S. Boll, IEEE Trans. Acoust., Speech and Signal Process., vol. ASSP-27, pp. 113-120, 1979.

[6] H. Hermansky and N. Morgan, IEEE Trans. Speech, Audio Process., vol. 2, pp. 578-589, 1994.

[7] 石塚健太郎，中谷智広，音講論，2-7-3, pp. 63-64, 秋季, 2005.

[8] M. Berouti, R. Schwartz, and J. Makhoul, Proc. ICASSP, pp. 208-211, 1979.

[9] A. Adami, L. Burget, S. Duponi, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas, Proc. Interspeech, pp. 21-24, 2002.

[10] B. S. Atal, J. Acoust. Soc. Am., vol. 55, pp. 1304-1312, 1974.

[11] C. P. Chen, K. Filali, and J. A. Bilmes, Proc. Interspeech, pp. 241-244, 2002.

[12] D. Macho, L. Mauuary, B. Noé, Y. M. Cheng, D. Ealey, D. Jouvet, H. Kelleher, D. Pearce, and F. Saadoun, Proc. Interspeech, pp. 17-20, 2002.