変調スペクトルによる音楽が付加された音声の自動検出の検討*

☆Pek Kimhuoch，荒井隆行（上智大），金寺登（石川高専）

1　はじめに

雑音・音楽環境下での音声区間検出(VAD)は，音声認識や音声情報処理をはじめとする分野において重要な役割を果たしており，近年では字幕翻訳の分野への応用も注目されている[1].

従来法では，ゼロ交差数や音声・非音声のエネルギー差[2]・線形予測に基づく方法[3]を特徴量として音声・非音声の判別が行われている．これらの手法は，比較的雑音の少ない環境を前提として自動音声区間検出を行っているが，雑音が多い環境下では十分な検出率が得られないという問題があった．この問題に対して，先行研究[4]では，音声に雑音が多く含まれる環境で自動的に音声部・非音声部を検出するための変調スペクトル手法を提案し，雑音環境下における音声・非音声の判別実験の結果改善が見られた．しかし，[4]では音楽が付加された音声の区間検出については検討されていなかった．本研究では，[4]の変調スペクトル手法にならい，5～15Hz の変調周波数帯域の変調指数を特徴量として，異なるジャンルの音楽を用いて実験を行った．

2　変調スペクトルによるVAD

変調スペクトルとは，入力音声特徴量の時間変化を周波数領域で表したものであり，その周波数領域は変調周波数と呼ばれている．先行研究[5-7]より，雑音環境下において変調周波数が 2 Hz 以下や 16 Hz 以上の変調スペクトル成分が音声認識性能を劣化させることが報告されている．また，先行研究[8]より，音声と音楽の識別に関しては，4 Hz 付近の変調エネルギーを利用することが報告された．これに対して，[4]では，雑音環境下で 5～15 Hz の変調周波数帯域を利用することで VAD 性能の改善を得ている．実環境における背景雑音には，雑音以外に音楽も含まれていることが多いことから，雑音を対象とした実験[4]と同じ変調周波数帯域を本報告では用いることにした．Fig. 1 に，音楽（ジャズ）中の音声区間検出のための特徴量を示す．短時間パワーでは，音声のパワー以外に音楽のパワーも検出してしまう．それに対して，変調スペクトルを用いた特徴量では，音声区間のみ大きな値を示している．このように，提案法は，音楽による影響を軽減できると予想される．

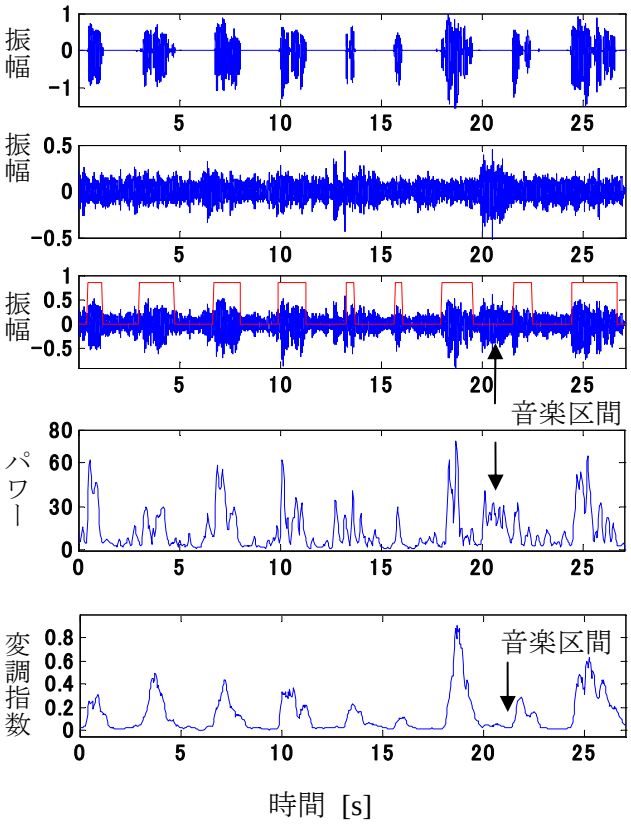

Fig. 1 Waveforms of a clean speech (top panel), additive music(2nd panel), clean speech combined with music (3rd panel, SNR=0 dB), power-based feature (4th panel), and modulation spectrum-based feature (5th panel). The red boxes are speech segments labeled by hand.

* Investigation on voice activity detection in music by using modulation spectrum, by PEK, Kimhuoch, ARAI, Takayuki (Sophia University) and KANEDERA, Noboru (Ishikawa National College of Technology).

3　実験データ及び評価方法

日本語の音声コーパス CENSREC-1-C[9]の数字音声を用いた．このコーパスのサンプリング周波数は 8kHz，量子化は 16bit，語彙は数字の 11 種類（1～9，ゼロ，まる），無音の計 12 種類である．CENSREC-1-C のクリーンデータに付加する音楽は RWC 研究用音楽データベースを用いた[10]．実験では，クリーンデータに異なる音楽を付加させて音声区間検出の結果を調べた．RWC の音楽ジャンルのデータベースからはクラシック，行進曲，ジャズ，ラテン（サンバ「曲目：Lovely Women」，レゲエ「曲目：Moon Struck」，タンゴ「曲目：Tango in Twilight」），ワールド（ブルース「曲目：Got'em Both」，フォーク「曲目：Grassy Dance」，カントリー「曲目：Bjc Fiddle Rag」，インド「曲目：Raga Charukesi- Drut Teentaal」，フラメンコ「曲目：Sevillanas」）と演歌を用いた．

入力音声データの周波数帯域を 250～2000 Hz に固定し，5～15 Hz の変調周波数帯域を用いた．フレーム長は 112.5 ms，フレームシフトはフレーム長の 1/3 を用いた．CENSREC-1-C の評価方法と同様，音声開始フレームと音声終了フレームにより決定される音声区間が 100 ms 未満の場合その区間を非音声区間として見なす．また、特徴量によって検出された発話区間の前後に 300 ms のマージンをとり発話区間として用いた．

実験の評価方法は発話単位での評価を行う．評価尺度は次式で定義される Correct rate (Corr, 発話区間検出正解率)，Accuracy (Acc, 発話区間検出正解精度)を用いた．

$$Corr = \frac{\text{正解発話区間検出数}}{\text{全発話数}}$$

$$Acc = \frac{\text{正解発話区間検出数} - \text{誤発話区間検出数}}{\text{全発話数}}$$

3.1 実験結果

Fig. 2 とFig. 3 は音声にラテン音楽を付加したときの結果を示す．発話区間検出正解率と発話区間検出正解精度をFig. 2 とFig. 3 に示す．比較対象に，パワーを特徴量とした結果を示した．また，Table 1 は変調スペクトルを特徴量としたときの異なる音楽ジャンルの結果を表わす．Fig. 2, 3 より雑音が増加につれて発話単位での正解率と正解精度が減少することがわかる．SNRが 0 dBまで，Corrが 90%以上でAccが 80%以上になっている．しかし，SNR が 0 dBより小さくなると付加する曲によって結果の下がり方は異なる．タンゴの曲は他の曲より正解率と正解精度が急に減少する傾向が見られる．また，変調スペクトルによる結果はパワーを特徴量とした結果と比べて，高SNR (20 dB, 10 dB) のときはほぼ同じ結果となっているが低SNR (0 dB, -2 dB, -5 dB, -7 dB) のときはCorrが 20%ほど，Accでは 40%以上の改善が見られる．タンゴの曲は他の曲より検出率が低いが，パワーと比べて改善が見られる．

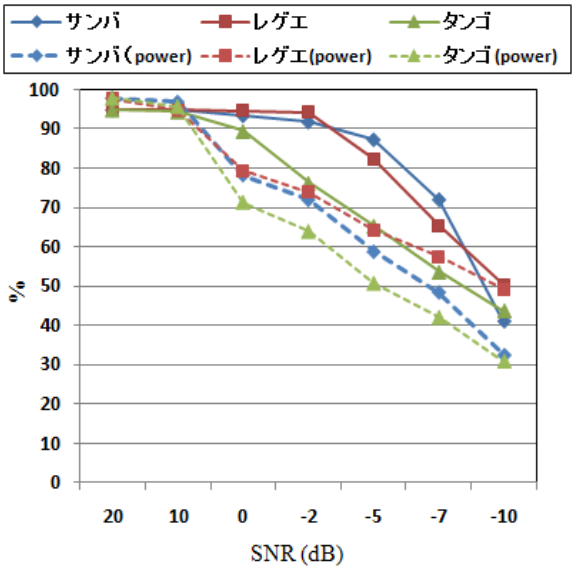

Fig. 2 Correct rate

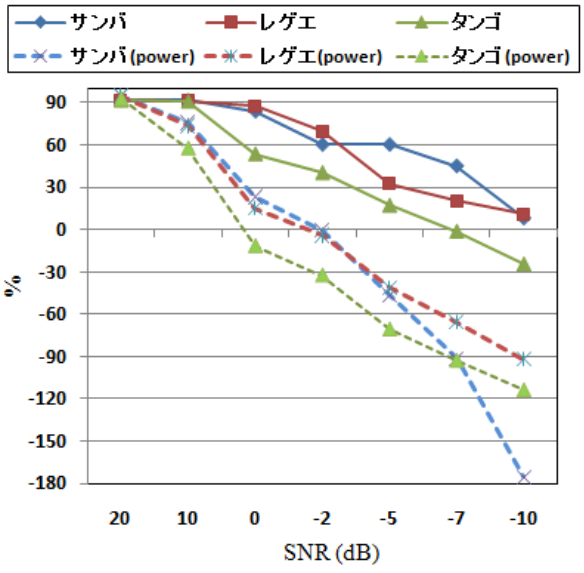

Fig. 3 Accuracy

Table 1 Results for different music types

		Correct Rate [%]					Accuracy [%]				
ジャンル	曲名 ＼ SNR (dB)	10	0	-2	-5	Overall	10	0	-2	-5	Overall
ジャズ	Wind Up	94.8	93.0	82.5	70.1	85.1	91.2	82.1	67.0	41.7	70.5
	Kitchen	94.5	91.9	83.5	63.2	83.3	90.8	71.3	58.3	30.7	62.8
	Azure	94.6	94.0	93.9	94.1	94.2	91.1	87.8	84.4	76.2	84.9
	Gypsy Eyes	94.5	92.4	69.0	50.5	76.6	90.9	81.1	42.9	7.7	55.7
	Wind Flower	95.0	94.5	94.2	81.2	91.2	91.6	78.7	66.5	48.4	71.3
ラテン	Lovely Women	94.9	93.5	91.9	87.4	91.9	91.6	83.9	60.2	60.3	74.0
	Moon Struck	94.8	94.5	94.3	82.4	91.5	91.3	87.6	69.8	32.3	70.2
	Tango in Twilight	94.5	89.5	76.3	65.4	81.4	90.9	53.7	40.5	17.4	50.6
クラシック	水上の音楽	94.9	91.8	81.8	64.4	83.2	91.6	82.2	60.8	21.8	64.1
	木星	94.5	77.0	57.5	42.4	67.8	90.8	55.1	31.4	8.2	46.4
行進曲	星条旗よ永遠なれ	94.9	85.9	67.0	57.5	76.3	91.4	61.1	34.4	17.0	51.0
ワールド	Got'em Both	94.9	94.7	94.6	93.8	94.5	91.6	63.0	62.6	58.4	68.9
	Grassy Dance	94.7	92.6	87.9	72.3	86.9	91.2	70.7	58.5	28.4	62.2
	Bjc Fiddle Rag	94.7	66.8	29.5	9.3	50.1	91.2	37.8	-1.2	-7.1	30.2
	Raga Charukesi	94.7	75.4	38.2	24.5	58.2	91.0	50.1	7.8	-9.8	34.8
	Sevillanas	95.0	93.5	91.5	74.5	88.6	91.8	66.6	60.5	51.3	67.6
邦楽	大漁船（演歌）	94.8	93.6	86.9	58.1	83.3	91.5	77.4	74.7	30.9	68.6

4　異なる音楽ジャンルによる考察

Table 1はSNRを変化させたときの異なる音楽ジャンルによるCorrとAccの結果を表す. この結果より，SNR=10 dBのとき使用されている全ての音楽ジャンルに対してCorrとAccの結果がほぼ同じである. しかし, 低SNR (0 dB, -2 dB, -5 dB)のときは音楽の曲によって結果が異なる．例として，ジャズのAzureとGypsy Eyes曲に関する結果については, Azureの場合はSNR=-5 dBでもCorr = 90%以上となり，Acc = 76%以上となっている．それに対して，SNR=-5 dB のとき，Gypsy Eyesにおける結果はCorr=50%, Acc = 7.7%となっている. また，これらの結果の中で最も検出率が低かったのはカントリー(全体のCorr=50.1%, Acc=30.2%)とインド (全体のCorr=58.2%, Acc=34.8 %)の音楽である.

音声のみ，インド，ジャズ，カントリー音楽の変調スペクトルを Fig. 4 に表す．音声の変調スペクトルは他の音楽の変調スペクトルより変調指数が高く，5～18 Hz 付近に局所的なピークがあるのに対して，インドとカントリー音楽における変調スペクトルのピークは変調周波数の広い範囲に分布し，複数のピークが存在している．一方，ジャズにおける変調スペクトルは 2～5 Hz 付近に局所的なピークがあるのに対して, 5 Hz 以上は一定の低い値を保てる．また，変調周波数の 5～15Hz 付近のインド・カントリーの変調指数はジャズに比べて高く見られる．よって，ジャズ音楽(Azure)における音声・音楽の検出率が他に使用している音楽よりも高かった理由は, 5～15 Hz における変調指数を特徴量として用いることだと考えられる．つまり，本手法は抽出した変調周波数帯域に残った信号を音声も音楽も同じに扱っているので音楽の振動するリズムが音声のエネルギー変化に似ている場合は音声・音楽の区別が困難になると言える.

Fig. 5 はクリーン音声・ジャズ・インド・カントリー音楽における変調スペクトルの特徴量の時間変化を表す．ジャズの変調スペクトル特徴量は音声と比べて，変調指数の値が低くなめらかに変化する．これに対して，インド・カントリーにおける特徴量は高く局所的に変化する．音楽区間で特徴量がしきい値より高くなると音声区間として見なされてしまう．特に，カントリーについては特徴量の時間変化のピークが多い（カントリー音楽も 5～18 Hz の変調周波数帯域における変調指数が高い）ため，音声・音楽の区別は他よりも困難である.

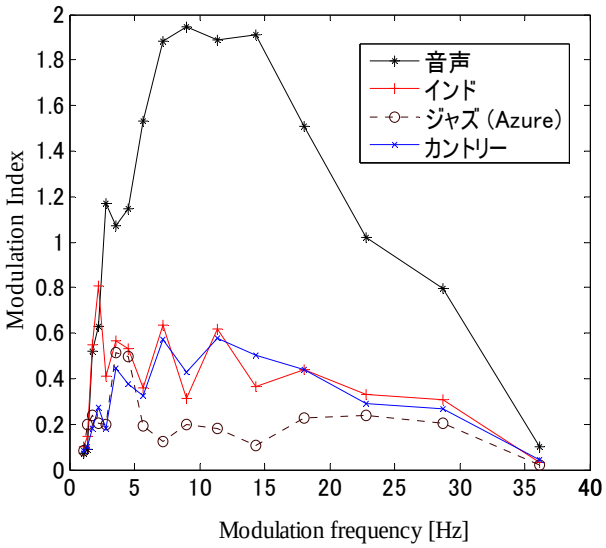

Fig. 4 Modulation spectrum of speech and music

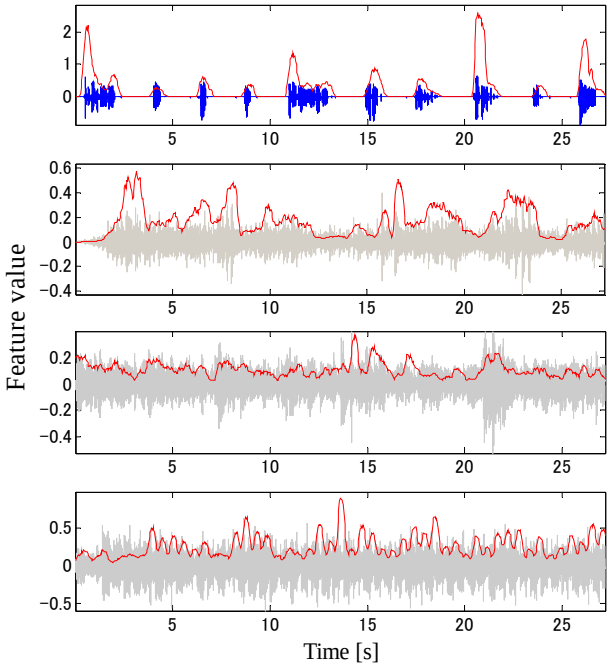

Fig. 5 Modulation spectrum-based feature of a clean speech (top panel), India music only (2nd panel), Jazz only (3rd panel) and country music only (4th panel). The red lines are the trajectories of the feature value.

5　おわりに

本研究では，変調スペクトルによる音声に音楽が重畳された場合の音声・非音声誤り率を調べた．変調スペクトル成分中で，5～15Hz の変調周波数帯域における変調指数を特徴量として用いた．この手法はパワーと比べて音声・音楽の判別結果の改善が見られた．さらに，異なる音楽ジャンルによる検出率を検討した．音楽ジャンルによる結果については，曲のリズムによって結果が異なる．

本実験では，雑音の実験[4]で利用している変調周波数成分を用いて実験を行った．しかし，音楽における他の有効な変調周波数成分については検討していない．今後の課題としては，音声に音楽が重畳されたとき，他の変調周波数成分を調べて，今回の実験結果と比較していくことやパワー以外の手法と比較していくことが重要である．

謝辞

この研究の一部は，文部科学省私立大学学術研究高度化推進事業　上智大学オープン・リサーチ・センター「人間情報科学研究プロジェクト」の支援を受けて行われた．また，評価のために CENSREC-1-C データベース，RWC 研究用音楽データベースを利用させて頂いた．

参考文献

[1] http://www.fujiyama1.com
[2] L. R. Rabiner et al, *BSTJ, 54 (2)*, pp. 287-315, 1975.
[3] 藤樫佑樹他，音講論，pp. 33-34，2005.
[4] K. Pek et al.，音講論, pp. 155-158, 2009.
[5] T. Arai et al., Proc. ICSLP, pp. 2490-2493, 1996.
[6] N. Kanedera et al. , Proc. Eurospeech, pp. 1079-1082, 1997.
[7] T. Arai et al., JASA,105(5), pp. 2783-2791, 1999.
[8] E. Scheirer et al., Proc. ICASSP, pp. 1331-1334, 1997.
[9] 北岡教英他，音講論, pp. 103-104, 2006.
[10] M. Goto et al., Proc. ISMIR , pp. 229-230, 2003.