2005－SLP－57（9）
2005／7／15

複数特徴の重み付き統合による雑音に頑健な発話区間検出

木田 祐介　　河原 達也

京都大学 情報学研究科 知能情報学専攻
〒 606-8501 京都市左京区吉田二本松町
e-mail: kida@ar.media.kyoto-u.ac.jp

あらまし　　複数の特徴を重み付き統合し、さらに重みを最適化することにより種々の雑音に頑健な発話区間検出 (VAD) を実現する手法を提案する。提案手法では、VAD の代表的な特徴である振幅レベル、ゼロ交差数、スペクトル情報、GMM 対数尤度の 4 つを統合的に用いる。これらの特徴の統合は、雑音環境に応じて最適な特徴を選択することを事実上包含しており、また統合重みを最適化することによりさらに検出性能の向上が期待できる。統合重みの最適化には最小誤り分類 (MCE) 学習を用いる。3 種類の雑音環境下での実験により、提案手法の雑音への頑健性を確認した。また重みの最適化が実際に検出性能を改善すること、また数回程度の発話で雑音環境に適応できることがわかった。

Voice Activity Detection based on Optimally Weighted Combination of Multiple Features

Yusuke Kida　　Tatsuya Kawahara

School of Informatics, Kyoto University,
Kyoto 606-8501, Japan

Abstract　　This paper presents a voice activity detection (VAD) scheme that is robust against noise, based on an optimally weighted combination of features. The scheme uses a weighted combination of four conventional VAD features: amplitude level, zero crossing rate, spectral information, and Gaussian mixture model likelihood. This combination in effect selects the optimal method depending on the noise condition. The weights for the combination are updated using minimum classification error (MCE) training. An experimental evaluation under three types of noisy environment demonstrated the noise robustness of our proposed method. Adapting the feature weights was shown to enhance the detection ability and to be possible using a few training utterance.

1　はじめに

現在の音声認識技術における最も重要な課題の一つに、雑音環境下での頑健な認識の実現が挙げられる。この問題を解決するための手法として、スペクトルサブトラクションや Wiener フィルターなどの雑音抑圧手法、MLLR や PMC による雑音へのモデル適応などのアプローチが知られている。これらに加えて、発話区間検出 (VAD) は雑音環境下での音声認識において非常に重要な要素技術である。音声区間が正しく検出されなければ、それに続く認識処理が成功する可能性はきわめて低くなる。近年 VAD に関する研究も盛んに行われ、これまでに様々なアルゴリズムが提案されている。しかし、未だ十分な性能が得られていないというのが現状である。さらに、それらのアルゴリズムの多くは性能が雑音条件に大きく依存してしまうという問題を含んでいる。

様々な種類の雑音に頑健な VAD を実現するために、本稿では複数の特徴を統合する枠組みを提案する。これは個別の手法を組み合わせて、対応できる雑音を全体として広げることを狙いとしており、雑音抑圧手法においても同様のアプローチが取られている [1]。本稿では、代表的な VAD 手法として知られている以下の 4 つの特徴を統合する。

- 振幅レベル
- ゼロ交差数
- スペクトル情報
- GMM 対数尤度

これらの特徴を重みを付けて統合し、それらの重みを最小誤り分類 (MCE) 学習によって最適化する。重み付き統合は、雑音環境に応じて最適な特徴を選択することを事実上包含している。また、統合重みの最適化によりさらに検出性能の向上が期待できる。その際、重みを最適化するまでの時間が重要になるが、本稿では 10 程度の発話で最適化を行うことを目指す。

本稿の構成は以下の通りである。第 2 章では、提案する VAD の枠組みと VAD に用いる各特徴、そして重みの最適化に用いる MCE 学習について述べる。第 3 章では実験条件及び結果を示し、第 4 章で結論と今後の課題について述べる。

2　VAD のための特徴の統合と重みの最適化

2.1　提案手法の枠組み

本稿で提案する VAD システムの概要を図 1 に示す。本システムではまず入力データをフレームに分割し、4 つの特徴量 (振幅レベル、ゼロ交差数、スペクトル情報、GMM 対数尤度) をそれぞれ計算する。それぞれの特徴量は図 1 では $f^{(1)}, \cdots, f^{(4)}$ と表記されている。そして、各特徴量に重み $w_1, \cdots, w_4$ を付けて統合する。ある時刻 t における入力フレーム $\boldsymbol{x}_t$ に対する統合スコア $F(\boldsymbol{x}_t)$ は以下の式で表される。

$$F(\boldsymbol{x_t}) = \sum_{k=1}^{4} w_k \cdot f^{(k)}(\boldsymbol{x_t}), \tag{1}$$

また、重み w_k は以下の制約条件を満たすものとする。

$$\sum_{k=1}^{4} w_k = 1, \tag{2}$$

$$w_k > 0, \tag{3}$$

ここで、重みの初期値は全て同値 (=0.25) とする。

統合スコアを得ると、次に音声か非音声かの判定を行う。判定には以下の二つの識別関数を利用する。

$$g_s(\boldsymbol{x}) = F(\boldsymbol{x}_t) - \theta, \tag{4}$$

$$g_n(\boldsymbol{x}) = \theta - F(\boldsymbol{x}_t), \tag{5}$$

ここで、θ は統合スコアの閾値である。$g_s(\boldsymbol{x}_t)$ が $g_n(\boldsymbol{x}_t)$ より大きい場合に $\boldsymbol{x}_t$ は音声フレームとみなされ、そうでない場合は非音声フレームとみなされる。なお、この判定は $F(\boldsymbol{x}_t)$ と θ との単純な比較によって実現することもできるが、MCE 学習では識別する各クラスタごとに識別関数を用意する必要があるため、ここでは二つの識別関数を用いている。この判定は入力フレームごとに行われるものとする。

重みの学習には、各発話における音声の開始フレームと終了フレームが人手により記述されたラベルファイルを用いる。重みの学習もフレーム単位で行われ、この枠組みを繰り返すことによって重みを最適化する。

2.2　検出に用いる特徴

2.2.1　振幅レベル

音声波形の振幅レベルは、VAD に用いられる最も基本的な特徴であり、様々な音声認識システムに実

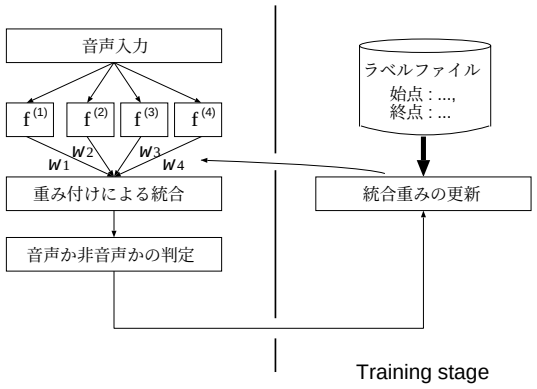

図 1: VAD の枠組み

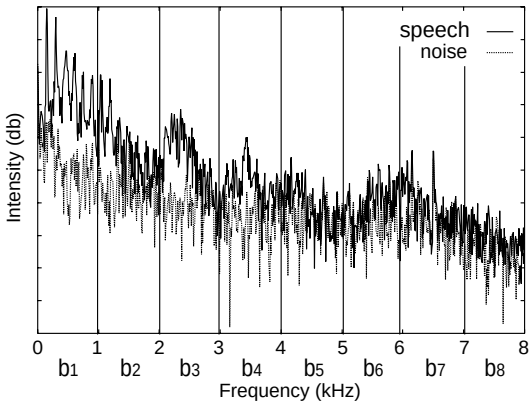

図 2: 音声 · 雑音のスペクトル例

装されている。t 番目のフレームに対する振幅レベル E_t は以下の式で求められる。

$$E_t = \log \sum_{n=1}^{N} s_n^2 \tag{6}$$

ここで、s_n はフレーム内のサンプル信号に長さ N のハミング窓をかけた値である。

本稿では雑音区間における特徴量が既知であるものとし、振幅レベルについては雑音レベルとの比を用いる。すなわち、統合に用いる特徴量 $f_t^{(1)}$ は

$$f_t^{(1)} = \frac{E_t}{E_n}, \tag{7}$$

となる。ここで、E_n は雑音区間での振幅レベル値である。

2.2.2 ゼロ交差数 (ZCR)

ゼロ交差数 (ZCR) は、一定時間内に信号レベルが 0 と交わる回数であり、音声区間ではこの値が大きくなることを利用して VAD に用いられる。ただし実際には 0 の代わりに一定のバイアス値を設定し、バイアスの範囲内での交差はカウントしないのが一般的である。特徴量 $f_t^{(2)}$ も振幅レベルと同様に雑音区間との比を用い、以下のように表される。

$$f_t^{(2)} = \frac{Z_t}{Z_n} \tag{8}$$

ここで Z_t は入力フレームの ZCR、Z_n は雑音区間での値である。

2.2.3 スペクトル情報

スペクトルから特徴を抽出して VAD に利用する研究は近年盛んに行われており、いくつかの手法が提案されている [2, 3]。音声と雑音のスペクトル例を図 2 に示す。ここでは、周波数領域をいくつかのチャネルに分割し、各チャネルごとに S/N 比を計算する。その上でチャネルごとの S/N 比の平均を取った値を特徴量とする。これは以下の式で表される。

$$f_t^{(3)} = \frac{1}{B} \sum_{b=1}^{B} 10 \log_{10} \frac{S_{bt}^2}{N_b^2}, \tag{9}$$

ここで、B はチャネルの数を表している。また、S_{bt} と N_b はそれぞれ入力フレーム及び雑音区間におけるチャネル b の平均パワーである。

2.2.4 GMM 対数尤度

ガウス混合分布 (GMM) は、統計的学習が容易なことから近年 VAD によく用いられている [4]。ここでは音声の GMM と雑音の GMM の対数尤度比を特徴として用いる。特徴量 $f_t^{(4)}$ は以下の式で示される。

$$f_t^{(4)} = \log(p(\boldsymbol{x}_t|\Theta_s)) - \log(p(\boldsymbol{x}_t|\Theta_n)), \tag{10}$$

ここで Θ_s,Θ_n は音声及び雑音 GMM のモデルパラメータセットである。

2.3 MCE 学習を用いた重み最適化

VAD システムを雑音環境に適応させるため、最小誤り分類 (MCE) 学習を用いて統合重みを最適化する。識別学習には、一般化確率的降下法 (GPD) [5] を用いる。

2.3.1 損失関数の定義

学習データ $\boldsymbol{x}_t$ に対する誤分類測度は以下のように表される。

$$d_k(\boldsymbol{x}_t) = -g_k(\boldsymbol{x}_t) + g_m(\boldsymbol{x}_t), \tag{11}$$

ここで k は正解クラスであり、音声 (s) か非音声 (n) に相当する。また、m は非正解クラスである。式 (11) が負のときには $\boldsymbol{x}_t$ が正しく分類されたことを示す。

次に、誤分類測度に 0,1 のステップ関数を近似するシグモイド関数を適用して、式 (12) より損失を定義する。

$$l_k(\boldsymbol{x}_t) = \big(1 + \exp(-\gamma \cdot d_k)\big)^{-1}, \tag{12}$$

ここで、γ はシグモイド関数の傾きを表す。確率的降下法に基づいて損失関数を最小にすることが識別学習の目標となる。

2.3.2 重みの最適化

本稿では、各特徴にかかる重みは常に 0 より大きくなければならない制約条件を設けている。MCE 学習による更新の過程において常にこの制約が満たされる (ことを保証する) ために、統合重み $\boldsymbol{w}$ を以下の新しいセット $\tilde{\boldsymbol{w}}$ に変換する。

$$\tilde{\boldsymbol{w}} = \{\tilde{w}_1, \tilde{w}_2, \ldots, \tilde{w}_L\}, \tag{13}$$

$$\tilde{w}_l = \log w_l, \tag{14}$$

$\tilde{\boldsymbol{w}}$ は、学習データが入力されるごとに以下のように更新される。ここで、ε_r は学習のステップを表し、データが入力されるたびに単調に減少していくものとする。

$$\tilde{\boldsymbol{w}}(t+1) = \tilde{\boldsymbol{w}}(t) - \varepsilon_r \nabla l_k(\boldsymbol{x}_t), \tag{15}$$

式 (15) の右辺最終項は、以下のように展開される。

$$\nabla_{\tilde{\boldsymbol{w}}} l_k(\boldsymbol{x}_t) = \frac{\partial l_k}{\partial d_k}\frac{\partial d_k}{\partial g_j} \cdot \nabla_{\tilde{\boldsymbol{w}}} g_j(\boldsymbol{x}_t), \tag{16}$$

ここで、$\frac{\partial l_k}{\partial d_k}$, $\frac{\partial d_k}{\partial g_j}$, $\nabla_{\tilde{\boldsymbol{w}}} g_j(\boldsymbol{x}_t)$ の要素 $\nabla_{\tilde{w}_l} g_j(\boldsymbol{x}_t)$ は以下のようになる。

$$\frac{\partial l_k}{\partial d_k} = \gamma l_k(1 - l_k), \tag{17}$$

$$\frac{\partial d_k}{\partial g_j} = \begin{cases} -1 & j = k \\ 1 & j \neq k \end{cases}, \tag{18}$$

$$\begin{aligned} \nabla_{\tilde{w}_l} g_j(\boldsymbol{x}_t) &= \frac{\partial}{\partial w_l}\left[\sum_{l=1}^{L} w_l F_l(\boldsymbol{x}_t)\right] \cdot \frac{\partial w_l}{\partial \tilde{w}_l} \\ &= w_l F_l(\boldsymbol{x}_t) \end{aligned} \tag{19}$$

$\tilde{\boldsymbol{w}}$ の学習が終了したら、$\tilde{\boldsymbol{w}}$ を $\boldsymbol{w}$ に逆変換する。

$$w_k = \frac{\exp \tilde{w}_k}{\sum_{l=1}^{L} \exp \tilde{w}_l} \tag{20}$$

ここで、式 (20) は制約条件 (2) を満たすための正規化処理も含んでいる。

3 実験と評価

3.1 タスク及び実験条件

本稿で提案した VAD システムの有効性を評価するために、雑音環境下における発話検出実験を行う。音声データは、10 人の話者の発話を防音室で収録したもの (16kHz,16bit) を用いる。1 人あたりの発話は 10 回で、各発話は 1~3 秒程度である。また、各発話の間には 3 秒程度のポーズが挿入されている。雑音の種類にはセンサールーム、工作機械、話し声の 3 種類を用意し、これを音声データに重畳することでテストデータを作成する。各雑音に対して重畳時の S/N 比を 10、15db とした 2 種類のデータを作成したので、テストデータのサンプルは計 600 (3 雑音 × 2 SNR × 10 人 × 10 発話) 発話となる。また重みの学習に用いるデータは、テストデータと同じ話者による別の 10 発話とする。本稿では式 (7)、(8) 及び (9) において、雑音の特徴量を計算する必要がある。今回は、当該データの音声が含まれていない最初の 1 秒間を用いてそれらを計算する。

次に、VAD に用いる特徴について述べる。フレーム長は振幅レベル及び ZCR においては 100 ミリ秒、GMM 対数尤度及びスペクトル情報については 25 ミリ秒とする。またフレーム周期は各特徴とも 10 ミリ秒とする。スペクトル情報の分割チャネル数は 20 とする。GMM には 32 混合で対角共分散行列のガウス分布を用い、その入力は 12 次元のメルケプストラム及びその一次差分 (Δ)、Δ-パワーとする。音声 GMM の学習には JNAS(新聞記事読み上げコーパス) の 304 人による約 32000 発話、雑音 GMM の学習にはセンサールーム、オフィス、廊下の 3 種類の雑音 (各 20 分程度) をそれぞれ用いる。ここで、センサールームのみが VAD の評価用データにも用いられている雑音である。

VAD の評価尺度には、フレームベースでの false alarm rate (FAR) 及び false rejection rate(FRR) を用いる。FAR は全非音声フレームにおいて誤って音声と認識されたフレームの割合、FRR は全音声フレームにおいて誤って非音声と認識されたフレームの割合をそれぞれ示す。

3.2 実験結果

6 パターンの雑音条件に対する実験結果を図 3〜8 に示す。それぞれの図は各特徴を単独で用いて VAD を行った結果と、統合して重みを最適化した提案手法の結果を表している。図中の 'Amplitude'、'ZCR'、'GMM'、'Spectrum' はそれぞれ振幅レベル、ゼロ交差数、GMM 対数尤度、スペクトル情報を、'Proposed' は提案手法を表している。また、図の横軸は FAR、縦軸は FRR に対応する。図中のプロットは識別関数の閾値に対応しており、閾値を変えながら実験を行うことによって図のようなオペレーション曲線を得た。

まず、単独の特徴について考察する。センサールーム雑音は雑音 GMM の作成に用いられたため GMM 対数尤度の結果が最もよくなることが期待されたが、実際には ZCR が最も高い性能を示した。また、工作機械ではスペクトル情報、話し声では GMM 対数尤度が最も高い VAD 性能を得た。これらの結果から、雑音環境に応じて最適な特徴が異なることがわかる。それに対して、提案手法は全ての雑音環境において、単独特徴を上回る結果を示した。これより、提案手法の有効性が示された。

次に、重み学習のために用いた音声データを評価する、いわゆるクローズド実験を行なった。センサールーム (S/N 比:10db) での実験結果を図 3 の 'Closed' に示す。図より、'Closed' と 'Proposed' の結果がほとんど同じであることがわかる。他の雑音条件についても同様の結果が得られた。これは、提案手法が発話の変動に対して頑健であることを表している。

また、重み適応の有効性を確かめるために、重みを最適化する前の状態 (すなわち全重みが等しい場合) で実験を行った結果と最適化後の結果を比較した。同時に、適応に用いる発話数を 1,5,10 と変化させて実験を行い、それに伴う性能の変化を調べた。各雑音を 10db で重畳したテストデータに対する実験結果を EER(Equal Error Rate) で表 1 に示す。EER は FAR と FRR が等しくなる点の値である。センサールームでは重み適応の前後で性能の変化がほとんど見られなかったが、工作機械と話し言葉では検出能力が改善され、全体としても雑音環境への適応の効果が見られる。また適応に用いた発話数において比較すると、1 発話より 10 発話の方が若干性能が向上したが、大きな違いは見られなかった。これより、1 回の発話でも重みの学習が有効であることがわかった。

表 1: 適応に用いる発話数を変えたときの EER (%)

	センサー	工作機械	話し声	平均
適応前	7.2	11.6	10.0	9.6
1 発話	7.1	9.9	9.6	8.9
5 発話	7.3	9.7	9.8	8.9
10 発話	7.2	9.7	9.4	8.8

4 おわりに

本稿では、複数の特徴の重み付き統合に基づく雑音に頑健な発話検出手法を提案した。統合する特徴には振幅レベル、ゼロ交差数、スペクトル情報、GMM 対数尤度の 4 つを用いた。3 種類の異なる雑音条件で実験を行った結果、提案手法はどの雑音環境に対しても個別の特徴より高い VAD 性能を実現した。また最小誤り分類学習による統合重みの最適化が、さらに検出性能を高めることを確認し、1 回の発話でも十分に学習が可能であることがわかった。

今後の課題としては、提案手法を用いた音声認識の実現及びその評価が挙げられる。

参考文献

[1] T.Yamada, J.Okada, K.Takeda, N.Kitaoka, M.Fujimoto, S.Kuroiwa, K.Yamamoto, T.Nishiura, M.Mizumachi, S.Nakamura, "Integration of noise reduction algorithms for Aurora2 task," *EUROSPEECH-2003*, pp.1769-1772, Sep.2003.

[2] J.Ramirez, J.C.Segura, C.Benitez, A.de la Torre, A.Rubio, "Voice Activity Detection with Noise Reduction and Long-Term Spectral Divergence Estimation,"*ICASSP-2004*, Vol.II, pp.1093-1096, May.2004.

[3] P.N.Garner, T.Fukada, Y.Komori, "A Differential Spectral Voice Activity Detector,"*ICASSP-2004*,Vol.I, pp.597-600, May.2004.

[4] A.Lee, K.Nakamura, R.Nishimura, H.Saruwatari, K.Shikano, "Noise Robust Real World Spoken Dialog System using GMM Based Rejec-

tion of Unintended Inputs,"*ICSLP-2004*, Vol.I, pp.173-176, Oct.2004.

[5] B.-H.Juang, S.Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Process.*, vol.40, no.12, pp.3043–3054, Dec. 1992.

[6] R.Nishimura, Y.Nishihara, R.Tsurumi, A.Lee, H.Saruwatari, K.Shikano,"Takemaru-kun: Speech-oriented information system for real world research platform,"*Int'l Workshop on Language Understanding and Agents for Real World Interaction*, pp.70-78, 2003.

[7] 井坂直人, 大須賀洋, 山田武志, 北脇信彦, 浅野太. 環境音モデルと HMM 合成を用いた音声区間検出法の音声認識への適用. 日本音響学会講演論文集, 1-4-19, Mar. 2003.

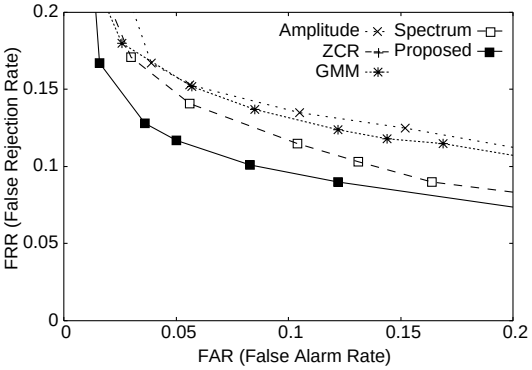

図 5: 工作機械:10db (ZCR は図の範囲外)

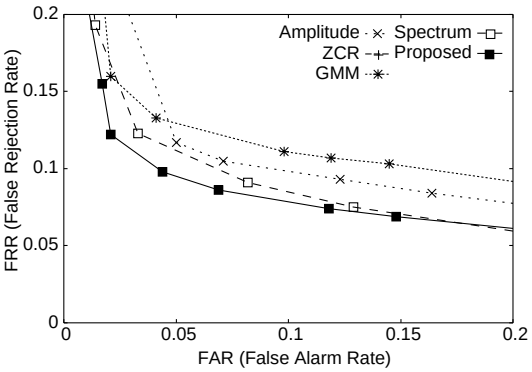

図 6: 工作機械:15db (ZCR は図の範囲外)

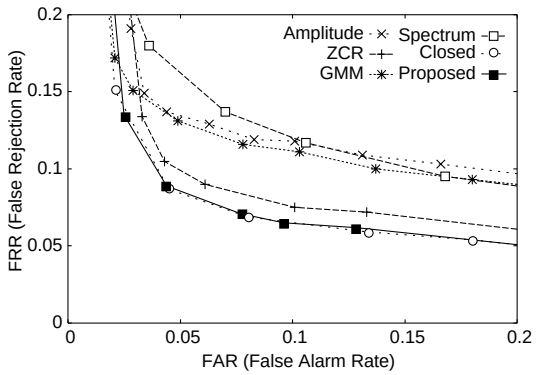

図 3: センサールーム:10db

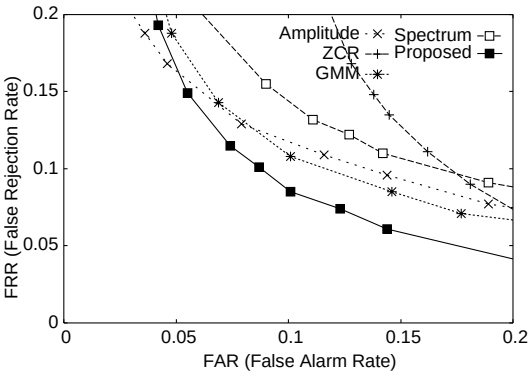

図 7: 話し声:10db

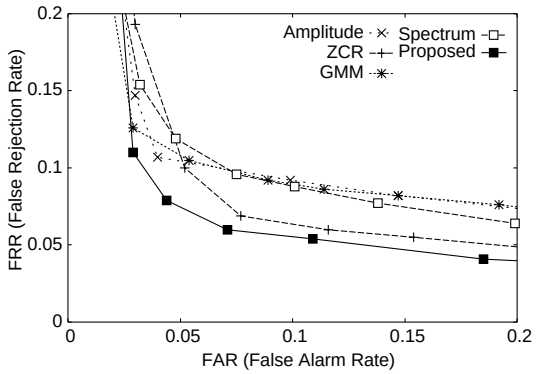

図 4: センサールーム:15db

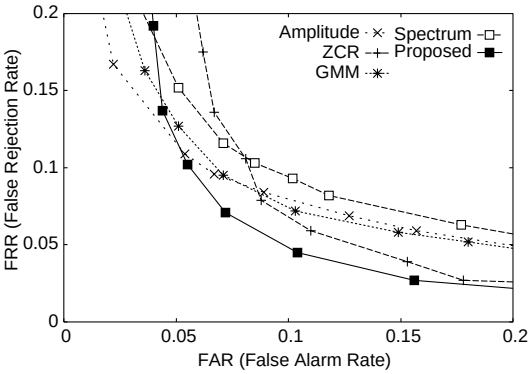

図 8: 話し声:15db