NMF に基づく音声と画像情報を用いた雑音下声質変換*

☆真坂健太 相原龍, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

声質変換は，入力された音声の言語情報を保ったまま，話者性や感情といった特定の情報のみを変換する技術である．音韻情報を維持しつつ話者情報を変換する “話者変換” [1] を目的として広く研究されてきたが，近年では，音声合成や音声認識における話者性の制御 [2] に用いられている他，感情情報を変換する “感情変換” [3, 4]，失われた話者情報を復元する “発話支援” [5] など多岐にわたって応用されている．本研究では，雑音環境下での声質変換など，これまでになかったタスクに対応可能な非負値行列因子分解（Non-negative Matrix Factorization : NMF）による声質変換 [6] を扱う．従来の NMF による声質変換では用いられていない唇画像特徴を声質変換に組み込むことで，変換精度の向上を目指した．

従来，声質変換においては統計的な手法が多く提案されてきた．なかでも混合正規分布モデル (Gaussian Mixture Model : GMM) を用いた手法 [1] はその精度のよさと汎用性から広く用いられており，多くの改良がされ続けられている．戸田ら [7] は従来の GMM を用いた声質変換法に動的特徴と Global Variance を導入することでより自然な音声として変換する手法を提案している．Helander ら [8] は従来手法における過適合の問題を回避するため，Partial Least Squares(PLS) 回帰分析を用いる手法を提案している．またこれらの手法では，入力話者と出力話者が同じテキストを発話して得られるパラレルデータが必要であるが，このパラレルデータを使用せずに声質変換を行うために，GMM の話者適応を行う手法 [9] や Eigen-Voice GMM (EV-GMM) [10, 11] などが提案されている．

しかし，これらの声質変換の従来手法のほとんどは学習・テストデータともにクリーン音声を用いており，雑音の重畳した入力音声に関する評価はされていない．入力音声に重畳した雑音は変換音声を生成する際の妨げとなり，その結果として変換される音声にも悪い影響が出ることは避けられない．よって雑音環境下を考慮した声質変換の手法の検討が必要であると言える．

我々はこれまで，従来の統計的手法とは異なる，スパース表現に基づく Exemplar-based な声質変換手法を提案してきた [6]．スパース表現に基づくアプローチは信号処理の分野において注目されており，音声信号処理の分野でも音声認識や音源分離，雑音抑圧などにおいて，その有効性が報告されている．このアプローチでは，与えられた信号は少量の学習サンプルや基底の線形結合で表現される．その後，目的音声の辞書に対する重みベクトルのみを取り出して用いることで，目的音声のみを分離する．Gemmeke ら [12] は雑音の重畳した音声を，クリーン音声辞書とノイズ辞書のスパース表現にし，クリーン音声辞書に対する重みを音声認識における Hidden Markov Model (HMM) の尤度算出に用いることで，雑音にロバストな音声認識を行う手法を提案している．

また，音声のみを用いた手法よりも画像と両方用いたマルチモーダルな手法が認識・変換においてよりよい結果をもたらすと言われている．駒井ら [13] は，雑音環境下において AAM を用いた発話認識手法を提案している．音声情報のみを用いた結果よりも画像情報をを取り入れたことでその有効性が示されている．

本研究では，スパースコーディングの代表的な手法として NMF [14] を用いる．我々の提案している声質変換手法では，従来の声質変換手法でも用いられていたパラレルデータから，入力話者の音声辞書（入力話者辞書）と出力話者の音声辞書（出力話者辞書）からなる同一発話内容のパラレル辞書を構築する．変換時には，入力音声を NMF によって，入力辞書に含まれる少量の基底からなるスパース表現にする．得られた入力辞書の基底毎の重み係数（アクティビティ）に基づいて，入力話者辞書の基底を出力辞書内の基底と置き換え，線形結合することで，出力話者の音声へと変換する．従来の声質変換のように統計的モデルを用いない Exemplar-based な手法であるため，過学習がおこりにくく，自然性の高い音声へと変換可能であると考えられる．

本稿では，雑音環境下に強い NMF 基づく声質変換に唇画像特徴を組み込んだ手法を提案する．ここでは入力音声の発話前後の非音声区間から雑音辞書を構築し，入力として与えられる雑音重畳音声を入力音声辞書と雑音辞書のスパースな表現にする．この入力音声と辞書から推定される重み行列のうち，音声辞書に関する重みのみを取り出し，出力話者の音声サンプルから構築した出力音声辞書との線形結合をとる．更に本手法では，入力話者の画像特徴から得られた唇画像辞書を導入することで変換精度をより向上させる．

以下，第 2 章でこれまでの NMF による声質変換手法を述べ，第 3 章で本稿の提案手法を説明する．第 4 章で従来の GMM・NMF による声質変換手法と比較し，第 5 章で本稿をまとめる．

2 NMF による声質変換

スパースコーディングの考え方において，与えられた信号は少量の学習サンプルや基底の線形結合で表現される．

$$\mathbf{x}_l \approx \sum_{j=1}^{J} \mathbf{a}_j h_{j,l} = \mathbf{A}\mathbf{h}_l \qquad (1)$$

$\mathbf{x}_l$ は観測信号の l 番目のフレームにおける D 次元の特徴量ベクトルを表す．$\mathbf{a}_j$ は j 番目の学習サンプル，あるいは基底を表し，$h_{j,l}$ はその結合重みを表す．本手法では学習サンプルそのものを基底 $\mathbf{a}_j$ とする．基底を並べた行列 $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_J]$ は “辞書” と呼び，重みを並べたベクトル $\mathbf{h}_l = [h_{1,l} \dots h_{J,l}]^T$ は “アクティビティ” と呼ぶ．このアクティビティベクトル $\mathbf{h}_l$ が

* Voice Conversion Based on Non-negative Matrix Factorization Using Audio and Visual Features in Noisy Environments. by Kenta Masaka, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki (Kobe University)

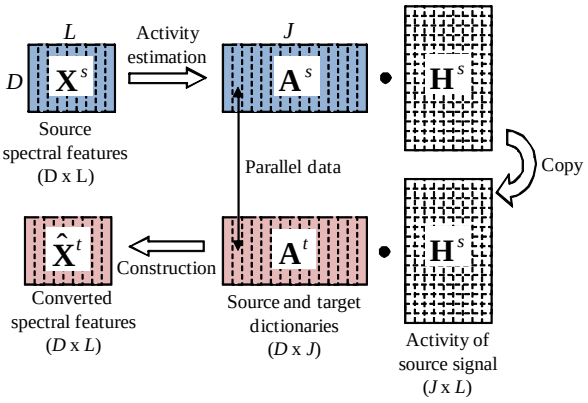

Fig. 1　Basic approach of NMF-based voice conversion

スパースであるとき，観測信号は重みが非ゼロである少量の基底ベクトルのみで表現されることになる．フレーム毎の特徴量ベクトルを並べて表現すると式 (1) は二つの行列の内積で表される．

$$\mathbf{X} \approx \mathbf{AH} \tag{2}$$

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_L], \quad \mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_L]. \tag{3}$$

ここで L はフレーム数を表す．

本手法の概要を Fig. 1 に示す．この手法では，パラレル辞書と呼ばれる入力話者音声辞書と出力話者音声辞書からなる辞書の対を用いる．この辞書の対は従来の声質変換法と同様，入力話者と出力話者による同一発話内容のパラレルデータに動的計画法 (DP) を適用することでフレーム間の対応を取った後，入力話者と出力話者の学習サンプルをそれぞれ並べて辞書化したものである．

このとき，仮に入力話者の音声と，それと同一発話の出力話者の音声をそれぞれ入力辞書と出力辞書のスパース表現にした場合，それぞれから得られるアクティビティ行列は互いに類似していると仮定できる．このことから，辞書行列がパラレルであれば，入力話者の辞書行列を用いて推定された入力特徴量のアクティビティは出力特徴量のアクティビティとして置き換え可能であると考えられる．以上の仮定に基づき，入力音声は入力話者辞書のスパース表現にし，得られたアクティビティ行列と出力話者辞書の内積をとることで，出力話者の音声へと変換する．本手法では，アクティビティ行列の推定にスパースコーディングの代表的手法である NMF を用いる．

3　唇画像特徴を用いた NMF 声質変換

これまでの NMF による声質変換法では，音声特徴のみを用いた変換手法となっていた．本稿では，唇画像特徴を組み込んだマルチモーダルな声質変換手法を提案する．これによってより雑音に頑健な変換となる．

3.1　辞書構成法

Fig. 2 は音声辞書，画像辞書の構成法を示したものである．従来の NMF による声質変換と同様にして各話者の同一発話によるパラレルデータから入力話者辞書 $\mathbf{A}^{sa}$ と出力話者辞書 $\mathbf{A}^{ta}$ を求める．本稿のテストデータには雑音が重畳しており，音声信号の分析合成ツールである STRAIGHT[17] ではその雑音を表現するのが難しという問題がある．従って，入力話者音声から構築する辞書内のサンプルは短時間フーリエ変換 (STFT) によって計算される振幅スペクトルとし，出力話者音声の辞書に関しては STRAIGHT 分析によって得られるスペクトルをサンプルとする．入力話者，出力話者ともに STRAIGHT 分析によって得られるメルケプストラムを用いて，フレーム間同期を取るための DP マッチングを行い，パラレルデータを作成する．画像辞書 $\mathbf{A}^{sv}$ に関して，フレームごとに得られる画像から特徴量を取り出し並べ，スプライン補間を行い音声フレームとの同期を取ったものを画像辞書とする．画像の特徴量として DCT（Discrete Cosine Transform）を用いる．DCT された画像からジグザグスキャン [15] を行い，低次 40 次元を負値を取らないように底上げしたものを画像特徴量とする．この画像辞書と音声辞書を結合したものを音声画像結合辞書 $\mathbf{A}^{s}$ とし，変換に用いるものとする．

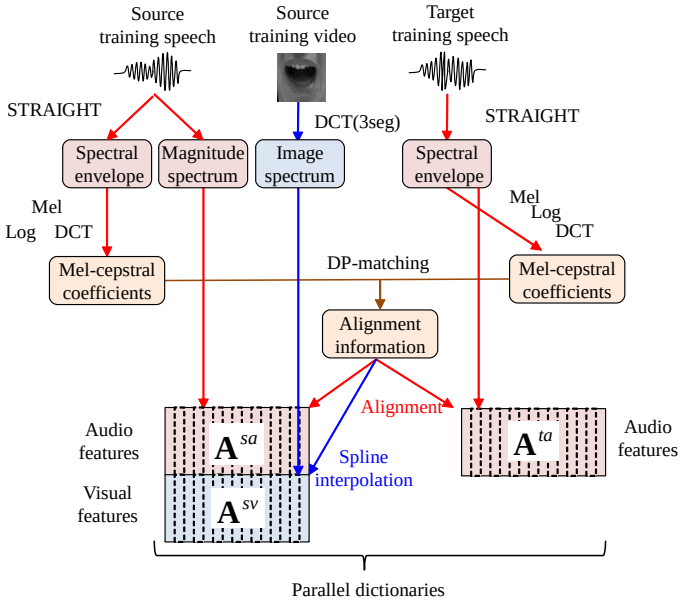

Fig. 2　Multimodal dictionary construction

3.2　変換手法

本稿で用いる変換手法は 2 段階に分かれている．1 段階目は従来の雑音除去 NMF [6] と同様，入力音声からノイズを除去する．入力話者の音声辞書に付随する雑音辞書は，雑音の重畳したテストデータの非音声区間のフレームから構築される．NMF による雑音除去手法において，観測信号のあるフレームは，クリーン音声から構築した辞書とノイズ辞書の非負の線形結合により近似される．

$$\begin{aligned}
\mathbf{x} &= \mathbf{x}^a + \mathbf{x}^n \\
&\approx \sum_{j=1}^{J} \mathbf{a}_j^{sa} h_j^a + \sum_{k=1}^{K} \mathbf{a}_k^{an} h_k^n \\
&= [\mathbf{A}^{sa}\mathbf{A}^{sn}] \begin{bmatrix} \mathbf{h}^a \\ \mathbf{h}^n \end{bmatrix} \quad s.t. \quad \mathbf{h}^a, \mathbf{h}^n \geq 0 \\
&= \mathbf{A}^{sa}\mathbf{h} \quad s.t. \quad \mathbf{h} \geq 0
\end{aligned} \tag{4}$$

$\mathbf{x}^a$ と $\mathbf{x}^n$ はそれぞれ入力話者のクリーン音声の振幅スペクトル，雑音の振幅スペクトルを表す．$\mathbf{A}^{sa}$, $\mathbf{A}^{sn}$, $\mathbf{h}^a$, $\mathbf{h}^n$ は入力話者の音声辞書，雑音の辞書，クリーン音声，雑音に対するそれぞれのアクティビティを表す．スペクトル，辞書はすべてフレーム毎に正規化されているものとする．スパース制約付き NMF において $\mathbf{h}$ を推定するためにコスト関数が以下のように設定されている．

$$d(\mathbf{x}, \mathbf{A}^{sa}\mathbf{h}) + ||\mathbf{1}^{(J\times 1)}\lambda^{\mathrm{T}}\mathbf{I}\mathbf{h}||_1 \quad s.t. \quad \mathbf{h} \geq 0. \tag{5}$$

第一項は $\mathbf{x}$ と $\mathbf{A}\mathbf{h}$ の Kullback-Leibler divergence である．第二項は $\mathbf{h}$ をスパースにするための L1 ノルム正則化項である．$\mathbf{1}$ はすべての要素が 1 の行列，$\mathbf{I}$ は単位行列を表す．$\lambda^T = [\lambda_1 \ldots \lambda_J \ldots \lambda_{J+K}]$ を調節することで，辞書内のサンプル毎に定義することができる．本稿ではクリーン音声辞書に関する制約重み $[\lambda_1 \ldots \lambda_J]$ を 0.1 に，雑音辞書に関する制約重み $[\lambda_{J+1} \ldots \lambda_{J+K}]$ を 0 に設定した．(5) 式を最小にするように以下の更新式に従いアクティビティ行列 $\mathbf{h}$ が推定される．

$$h_j \leftarrow h_j \frac{\sum_d^D (A^{sa}_{d,j} x_d / (\mathbf{A}^{sa}\mathbf{h})_d)}{1+\lambda_j} \tag{6}$$

D は辞書 $\mathbf{A}^{sa}$ の次元数を表す．このアクティビティ行列 $\mathbf{h}$ から入力話者辞書に関するアクティビティ $\mathbf{h}^a$ のみを取り出し，入力話者の辞書との内積を取ることで，雑音除去された入力話者の振幅スペクトル $\hat{\mathbf{x}}^a$ を得る．

$$\hat{\mathbf{x}}^a = \mathbf{A}^{sa}\mathbf{h}^a \tag{7}$$

2 段階目は，1 段階目で得られた振幅スペクトルに画像特徴量を結合する．これを入力として，3.1 章で定義した音声画像結合辞書 $\mathbf{A}^s$ からアクティビティ行列を推定する．ここで，入力，辞書にそれぞれ音声と画像の重み，α と β を掛けることにする．このとき，音声と画像を結合した特徴量から得られる新たな $\mathbf{h}^{av}$ は以下のコスト関数を最小にすることで推定でき，そのコスト関数と更新式は以下のようになる．

$$\alpha d(\hat{\mathbf{x}}^a, \mathbf{A}^{sa}\mathbf{h}^{av}) + \beta d(\mathbf{x}^v, \mathbf{A}^{sv}\mathbf{h}^{av}) + ||\mathbf{1}^{(J\times 1)}\lambda^{\mathrm{T}}\mathbf{I}\mathbf{h}^{av}||_1 \quad s.t. \quad \mathbf{h}^{av} \geq 0. \tag{8}$$

$$h^{av}_j \leftarrow h^{av}_j \frac{\sum_d^D f_d + \sum_e^E g_e}{\alpha+\beta+\lambda_j} \tag{9}$$

$$\mathbf{f}_d = \alpha A^{sa}_{d,j} \alpha \hat{x}^a_d / (\alpha \mathbf{A}^{sa}\mathbf{h}^{av})_d \tag{10}$$

$$\mathbf{g}_e = \beta A^{sv}_{e,j} \beta x^v_e / (\beta \mathbf{A}^{sv}\mathbf{h}^{av})_e \tag{11}$$

ただし $\hat{\mathbf{x}}^a$,$\mathbf{x}^v$,$\mathbf{A}^{sa}$,$\mathbf{A}^{sv}$ はそれぞれ入力音声，入力画像，音声辞書，画像辞書を表す．推定されたアクティビティ行列と出力話者辞書 $\mathbf{A}^{ta}$ の内積を取り，NMF 変換後のスペクトル $\hat{\mathbf{x}}^t$ を得る．

$$\hat{\mathbf{x}}^t = \mathbf{A}^{ta}\mathbf{h}^{av} \tag{12}$$

4 評価実験

4.1 実験条件

本実験では従来の GMM を用いた手法と，音声特徴のみを用いた NMF による手法を比較手法として実験を行った．M2TINIT 研究用マルチモーダル音声データベース [16] より，男性話者 1 名の音声を入力話者音声に，ATR 研究用日本語音声データベース [17] より，女性話者 1 名の音声を出力話者音声として用いた．サンプリング周波数は 16kHz である．音素バランス 50 文からパラレルデータを作成し，NMF におけるパラレル辞書の構築，従来手法における GMM の学習にそれぞれ用いた．各話者の辞書に含まれるサンプルの数は 88,275 である．テストデータとなる文章それぞれに雑音信号を加算した．雑音信号は CENSREC-1-C データベース [18] に含まれる食堂内で収録された音声の無音声部分の雑音を用いた．雑音信号の SNR は 0 dB と 10 dB とした．入力話者の音声辞書，及び入力音声に用いる振幅スペクトルの次元数は 512 ，出力話者の音声辞書，及び出力音声の生成に用いる STRAIGHT スペクトルの次元数は 1,025 である．アクティビティ行列の推定値の更新回数は 300 とした．比較手法である GMM では STRAIGHT スペクトルから計算される線形ケプストラム 24 次元を特徴量とし，混合数は 64 とした．動画のフレームレートは 29.97 fps で，画像のサイズは 720 x 840 となっている．音声フレームと画像フレームの同期を取るためにスプライン補間を行い，セグメント特徴量を導入し，前後 3 フレーム分を画像特徴として画像辞書を構築した．また，画像と音声の重みについては音声に対する重み α は 1 に固定し，画像に対する重み β を 0.01 , 0.1, 1 , 10 , 100 と変えて評価を行った．

4.2 実験結果・考察

各手法における変換音声の SDIR（Spectral Distortion Improvement Ratio）を Fig. 3 と Fig. 4 に示す．SDIR は以下の式で表される．

$$\mathrm{SDIR[dB]} = 10\log_{10}\frac{\sum_d |\mathbf{X}^t(d) - \mathbf{X}^s(d)|^2}{\sum_d |\mathbf{X}^t(d) - \hat{\mathbf{X}}^t(d)|^2} \tag{13}$$

ただし X^s,X^t,$\hat{X}^t$ はそれぞれ入力話者のスペクトル，出力話者のスペクトル，変換後のスペクトルを表す．スペクトルはそれぞれ次元ごとに正規化されている．Fig. 3,Fig. 4 より，0 dB ， 10 dB ともに画像特徴を用いた変換精度が，二つの従来手法より向上していることが分かる．また，重み β に関しては 10 のとき精度が一番良くなっている．β が 0 と 10 のときの SDIR の差と比べると，0 dB のときは 0.26，10 dB のときは 0.16 と重畳雑音の SN 比が小さいほうが，画像を入れたときの効果が大きいということも分かる．

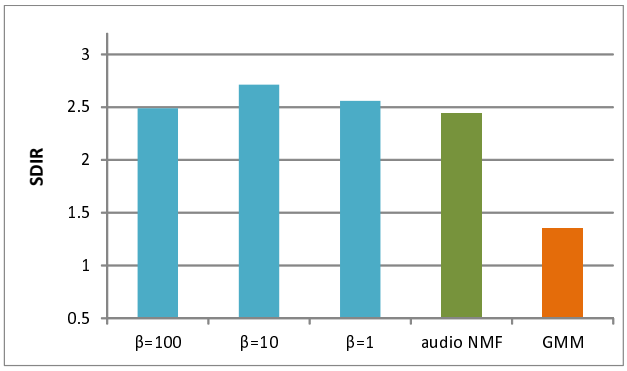

Fig. 3　Spectral Distortion Improvement Ratio calculated from converted speech using each method and weigth (SNR = 0 dB)

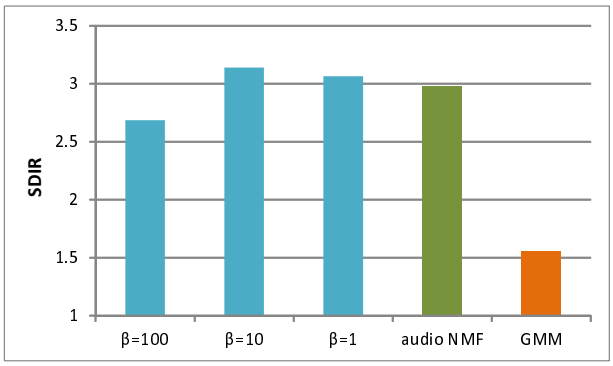

Fig. 4　Spectral Distortion Improvement Ratio calculated from converted speech using each method and weigth (SNR = 10 dB)

5　おわりに

本稿では，これまで提案してきたNMFに基づく声質変換法において，画像特徴を導入した．

これにより，音声特徴のみを用いた変換よりもさらに雑音に頑健な変換を行うことを可能となり変換精度が向上した．評価実験を行い，従来の統計的モデルを用いた声質変換法や音声特徴のみを用いたNMFよりも高い精度で変換できることを示した．

今後は他の画像特徴量での比較や，より精度の高い変換手法の改良を進めていく．

謝辞　本研究はJST A-STEPの成果である．

参考文献

[1] Y. Stylianou *et al.*, “Continuous probabilistic transform for voice conversion,” IEEE Trans. Speech and Audio Processing, vol. 6, no. 2, pp. 131–142, 1998.

[2] A. Kain and M. W. Macon, “Spectral voice conversion for text-to-speech synthesis,” in ICASSP, vol. 1, pp. 285-288, 1998.

[3] C. Veaux and X. Robet, “Intonation conversion from neutral to expressive speech,” in Interspeech, pp. 2765–2768, 2011.

[4] R. Aihara *et al.*, “GMM-based emotional voice conversion using spectrum and prosody features,” American Journal of Signal Processing, vol. 2, no. 5, 2012.

[5] K. Nakamura *et al.*, “Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech,” Speech Communication, vol. 54, no. 1, pp. 134–146, 2012.

[6] R. Takashima *et al.*, “Exemplar-based voice conversion in noisy environment,” in SLT, pp. 313–317, 2012.

[7] T. Toda *et al.*, “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 8, pp. 2222–2235, 2007.

[8] E. Helander *et al.*, “Voice conversion using partial least squares regression,” IEEE Trans. Audio, Speech, Lang. Process., vol. 18, Issue:5, pp. 912-921, 2010.

[9] C. H. Lee and C. H. Wu, “Map-based adaptation for speech conversion using adaptation data selection and non-parallel training,” in Interspeech, pp. 2254–2257, 2006.

[10] T. Toda *et al.*, “Eigenvoice conversion based on Gaussian mixture model,” in Interspeech, pp. 2446–2449, 2006.

[11] D. Saito *et al.*, “One-to-many voice conversion based on tensor representation of speaker space,” in Interspeech, pp. 653–656, 2011.

[12] J. F. Gemmeke and T. Virtanen, “Noise robust exemplar-based connected digit recognition,” in ICASSP, pp. 4546–4549, 2010.

[13] Y. Komai *et al.*, “Robust aam-based audio-visual speech recognition against face direction changes,” ACM Multimedia, pp. 1161–1164, 2012.

[14] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” Neural Information Processing System, pp. 556–562, 2001.

[15] J. Almajai and B. Milner, “Using audio-visual features for robust voice activity detection in clean and noisy speech,” in EUSIPCO, 2008.

[16] S. Sako *et al.*, “HMM-based text-to-audio-visual speech synthesis - image-based approach,” ICSLP, vol. III, pp. 25-28, 2000.

[17] A. Kurematsu *et al.*, “ATR japanese speech database as a tool of speech recognition and synthesis,” Speech Communication, vol. 9, pp. 357–363, 1990.

[18] N. Kitaoka *et al.*, “CENSREC-1-C: An evaluation framework for voice activity detection under noisy environments,” Acoustical Science and Technology, Vol. 30 (2009), 2009.