

音声生成過程と信号観測過程のモデルに基づく マルチチャンネル音声強調

中谷 智広^{1,a)}

概要: 日常生活の中で音声をマイクロホンで収録すると、多くの場合、目的音声以外の雑音（TV音、雑踏、他の話者の声）が混ざってしまう。このため、コンピュータによる音声認識が困難になる。これに対し、本稿では、雑音が混ざった観測信号から目的音声のみを抽出（＝音声強調）し、コンピュータによる音声認識を可能にする技術について議論する。まず、既存の様々な音声強調技術の基礎となる生成モデルアプローチの考え方について述べ、その考え方に基づき各技術の特徴を概観する。次に、生成モデルアプローチの枠組みの中で、複数マイクロホン処理に基づく雑音抑圧法と、音声認識に適したスペクトル推定法を統合することで、大幅な音声認識性能の改善が得られることを紹介する。

1. はじめに

音声は、多くの人にとって、最も身近で利便性の高いコミュニケーション手段である。日常生活の中で人が話す音声をコンピュータが理解できれば、人の活動を支援する技術の適用領域がさらに広がると期待される。近年、スマートホン等を用いた音声認識インタフェースが利用されるようになってきた。しかし、これらのインタフェースでは、利用者が、比較的静かな環境で話す必要があるなど、まだ、その適用領域に大きな制限がある。

一方、人は、健聴者であれば、目的音声以外の雑音（TV音、雑踏、他の話者の声）が聞こえている状況でも、目的音声を聞き分け、理解することができる。これは、日常環境で、人と人が円滑な音声コミュニケーションを行うために重要な能力といえる。同じ状況で、コンピュータが、人に快適な音声インタフェースを提供するためには、同様に高度な聞き分け能力が必要になると考えられる。

本稿では、雑音が含まれる観測信号から目的音声を抽出（＝音声強調）し、高い音声認識性能を得るための技術について議論する。まず、既存の様々な音声強調技術の基礎となる生成モデルアプローチ [1] について述べるとともに、各音声強調技術の特徴を概観する。次に、生成モデルアプローチの枠組みの中で、高い音声認識性能を実現する音声強調法を構築する目的で、複数マイクロホン処理に基づく雑音抑圧法と、音声認識に適したスペクトル推定法を統合

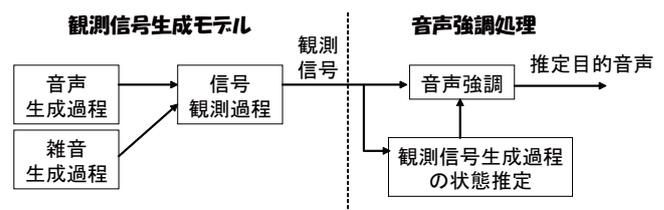


Fig. 1 Generative model based speech enhancement

する方法を紹介する [2].

2. 生成モデルアプローチに基づく音声強調

音声や雑音がどのような物理的・統計的制約に従い各音源から生成され（音源生成過程）、どのような変形を受けて収録・混合されるか（信号観測過程）がわかっている場合、マイクロホンで収録される音（＝観測信号）がどうなるかを予測することができる。本稿では、この過程を観測信号生成モデルと呼ぶ（図1左参照）。一方、観測信号生成モデルが、どのような制約に従うかがわかっている時に、与えられた観測信号に対し、観測信号生成モデル中の各過程がとりうる尤もらしい状態を推定し（図1中の観測信号生成過程の状態推定）、さらに、推定した状態に基づき観測信号中の目的音声を推定（＝音声強調）するアプローチを、本稿では、生成モデルアプローチと呼ぶ。

例えば、いま、仮に、目的音声や雑音の特徴量 $x^{(s)}$ の生成過程（＝音源生成過程）が、ある確率密度関数 $p(x^{(s)}; \theta^{(s)})$ に従うと仮定できるとする。 s は目的音声 ($s = 1$)、もしくは雑音 ($s = 2$) のいずれかを表す番号とし、 $\theta^{(s)}$ は、確率密度関数の形状を決定する状態パラメータとする。さらに、各音源から出た音が混ざりあってマイクロホンで収録

¹ 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
NTT Communication Science Labs., Kyoto 619-0237, Japan

^{a)} nakatani.tomohiro@lab.ntt.co.jp

され、観測信号の特微量 y として得られる過程 (=信号観測過程) が、ある確率密度関数 $p(y|x^{(1)}, x^{(2)}; \theta^{(o)})$ に従うと仮定できるとする。 $\theta^{(o)}$ は、信号観測モデルを規定する状態パラメータとする。すると、観測信号の確率密度関数 (=観測信号生成モデル) は、以下のようにモデル化できる。

$$p(y; \Theta) = \int p(y|x^{(1)}, x^{(2)}; \Theta) \prod_s p(x^{(s)}; \Theta) dx^{(s)}. \quad (1)$$

ここで、 $\Theta = \{\theta^{(1)}, \theta^{(2)}, \theta^{(o)}\}$ は、全状態パラメータの集合を表す。このモデルでは、 Θ が与えられれば、 $p(y; \Theta)$ が定まり、観測信号の挙動が予測できることになる。一方、一旦、上記のモデルが定まると、今度は、逆に、観測信号の特微量 y が与えられた場合に、 y を観測信号として出力しうる状態パラメータ Θ を推定することができるであろう。これには、例えば、 $p(y; \Theta)$ を最大にするという基準を用いることができる。また、さらに、状態パラメータの推定値 $\hat{\Theta}$ が与えられれば、例えば、最小自乗誤差基準などにより、各構成音の特微量を以下のように推定することができるであろう。

$$\hat{x}^{(s)} = \int x^{(s)} p(x^{(s)}|y, \hat{\Theta}) dx^{(s)} \quad (2)$$

これが生成モデルアプローチの基本的な考え方である。上記では、最尤法に基づく例を示したが、他の統計的推論の枠組みでも同様の定式化は可能である。

2.1 生成モデルアプローチの基本課題

上記に示した生成モデルアプローチの利点は、1) 信号の様々な統計的/物理的性質を各過程のモデルとして導入することで、複雑で多様な観測信号生成モデルを比較的容易に構築できること、また、2) 構築した観測信号生成モデルから、各過程の状態推定を行うための最適化基準が、そのまま自然に定まることである。

一方、本アプローチの問題は、多くの場合、各過程の状態推定が非線形最適化問題となるため、必ずしも実現が容易ではないことである。

このため、生成モデルアプローチの基本課題は、音声強調の目的にとって重要な特微量を利用しつつ、全体として、効率的な状態推定が可能な系を見つけることである。

2.2 生成モデルアプローチに基づく音声強調の例

近年、生成モデルアプローチに基づく効果的な音声強調法が、数多く提案されている。それらは、主に、音の空間的特徴に基づくモデルを用いる方法と、スペクトル特徴に基づくモデルを用いる方法に大別できる。

音の空間的特徴に基づくモデルを用いる方法では、複数マイクロホンにより収録した観測信号に対し、各音源から各マイクロホンまでの音響伝達特性の違いをモデル化することで、各構成音を区別できるようにし、音声強調を実現

する。この考え方に基づき、数多くのブラインド音源分離法 (BSS) が提案されている [3], [4], [5], [6]。これらの手法は、比較的高い雑音抑圧性能が達成できる一方、音声スペクトル形状の推定の正確さについては全く考慮されないという問題がある。

一方、音のスペクトル特徴に基づくモデルを用いる方法では、目的音声や雑音を取りうるスペクトルの形状をモデル化することで音声と雑音を区別し、音声強調を実現する。スペクトル形状のモデル化には、信号のパワースペクトルを非負値行列分解を用いてモデル化する方法 [8], [9]、メル対数スペクトルのように音声認識で用いられる特微量を混合ガウス分布などでモデル化する方法 [7], [10], [11]、音声事例に基づく方法 [12], [13], [14] など、多数のものが提案されている。これらの方法の多くは、目的音声のスペクトルがとりうる値を事前学習したモデルを用いるため、推定されるスペクトルが、ある程度目的音声に近いものになるという特長がある。しかし、音の到来方向など、雑音抑圧に効果的な手がかりを利用できないため、必ずしも高い雑音抑圧性能が達成できるとは限らないという問題がある。

3. 音声認識のための複数マイクロホン音声強調

上述のように、音の空間的特徴とスペクトル特徴は、観測信号中の目的音声を区別する上で、相補的な役割を果たしうる特徴である。両方の特徴に関するモデルをうまく統合した音声強調法が構築できれば、より高精度な音声強調が実現できることが期待される [15], [16]。筆者らも、その一環として、音声認識のための複数マイクロホン音声強調法を提案している [2]。

筆者らの提案法は、特に、音声認識の高精度化を目的としており、高いスペクトル推定精度と雑音抑圧精度を両立するために、上述の二つの特微量に対して、以下のモデルを採用している。

- スペクトル特徴: 各音源のメル周波数ケプストラム係数 (MFCC) を混合ガウス分布でモデル化した音源生成過程と、Factorial モデルでモデル化した信号観測過程を用いる。
- 空間的特徴: 各音源のステアリングベクトルを複素ワトソン分布でモデル化した音源生成過程と、混合複素ワトソン分布でモデル化した信号観測過程を用いる。

前者は、モノラル音声分離・認識チャレンジ [18] において、トップスコアを達成したモデル [17] の拡張に相当する。また、後者は、SiSEC 音源分離キャンペーン [19] において、高い音源分離性能を達成したモデル [5] と同一のものである。提案法では、これら二つのモデルを統合した観測信号生成モデルを考案するとともに、期待値最大化アルゴリズムを用いて、効率的な状態パラメータ推定法を構築している。その結果、提案法を他の最先端の音声認識技術

とを組合せて、生活雑音環境下での音声コマンド認識 [20]、および、複数人会話の音声認識 [21] に適用することで、大幅な音声認識性能改善が得られることが示されている。

4. 今後の課題

音の空間的特徴とスペクトル特徴を統合的に利用する音声強調法は、まだ、発展途上の研究課題である。今後、更なる性能改善が期待される [22]。また、日常環境での音声強調を実現するうえで、目的話者の数が動的に変化する環境を扱えるようにすることは重要である。近年、生成モデルアプローチの観点からも、これを解決する試みがなされている [23]。

謝辞 本稿で紹介した筆者らの研究は、NTT コミュニケーション科学基礎研究所信号処理研究グループのメンバーとの共同成果である。本稿執筆にあたり、同メンバーには多大な協力をいただいた。

参考文献

- [1] 亀岡弘和: 生成モデルアプローチによる音響信号処理, 奈良先端音楽技術大学院大学ゼミナール I 講演, 入手先 (<http://www.brl.ntt.co.jp/people/kameoka/publications/Kameoka2012NAISTseminar12slide.pdf>), 12 月 18 日, (2012).
- [2] Nakatani, T., Yoshioka, T., Araki, S., Delcroix, M., and Fujimoto, M.: Logmax observation model with MFCC-based spectral prior for reduction of highly nonstationary ambient noise," *IEEE ICASSP'12*, pp. 4029–4033 (2012).
- [3] Hyvärinen, A., Karhunen, J., and Oja, E.: *Independent Component Analysis*, John Wiley & Sons (2001).
- [4] Duong, N.Q.K., Vincent, E., and Gribonval, R.: Underdetermined reverberant audio source separation using a full-rank spatial covariance model, *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, Sep. (2010).
- [5] Sawada, H., Araki, S., and Makino, S.: Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment, *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 516–527 (2011).
- [6] Yoshioka, T., Nakatani, T., Miyoshi, M., and Okuno, H.G.: Blind separation and dereverberation of speech mixtures by joint optimization, *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 1, pp. 69–84, Jan. (2011).
- [7] Fujimoto, M., Watanabe, S., and Nakatani, T.: Frame-wise model re-estimation method based on Gaussian pruning with weight normalization for noise robust voice activity detection, *Speech communication*, vol. 54, no. 2, pp. 229–244, Feb. (2012).
- [8] Mysore, G., Smaragdakis, P., and Raj, B.: Non-negative hidden Markov modeling of audio with application to source separation, *Proc. LVA/ICA* (2010).
- [9] 亀岡弘和: 非負値行列因子分解とその音響信号処理応用 (招待講演), 電子情報通信学会技術報告, vol. 112, no. 347, EA2012-118, pp. 53–58, Dec. (2012).
- [10] Moreno, P.J., Raj, B., and Stern, R.M.: A vector Taylor series approach for environment-independent speech

- recognition, *Proc. IEEE ICASSP'96*, vol. 2, pp. 733–736 (1996).
- [11] Roweis, S.T.: Factorial models and refiltering for speech separation and denoising, *Proc. Interspeech'03*, pp. 1009–1012 (2003).
- [12] Ming, J., Srinivasan, R., and Crookes, D.: A corpus-based approach to speech enhancement from nonstationary noise, *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 4, pp. 822–836 (2011).
- [13] Kinoshita, K., Souden, M., Delcroix, M., and Nakatani, T.: Single channel dereverberation using example-based speech enhancement with uncertainty decoding technique, *Proc. Interspeech'11*, pp. 197–200 (2011).
- [14] Sainath, T.N., Ramabhadran, B., Nahamoo, D., Kanevsky, D., Van Compernelle, D., Demuynck, K., Gemmeke, J.F., Bellegarda, J.R., and Sundaram, S.: Exemplar-based processing for speech recognition, *IEEE Signal Process. Magazine*, vol. 29, no. 6, pp. 98–113, Nov. (2012).
- [15] Ozerov, A. and Févotte, C.: Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis, *Neural Computation*, vol. 21, no. 3, pp. 793–830 (2009).
- [16] Sawada, H., Kameoka, H., Araki, S., and Ueda, N.: Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization, *Proc. ICASSP'12*, pp. 261–264 (2012).
- [17] Rennie, S.J., Hershey, J.R., and Olsen, P.A.: Single-channel multitalker speech recognition, *IEEE Signal Process. Magazine*, vol. 27, no. 6, pp. 66–80 (2010).
- [18] Cooke, M., Hershey, J., and Rennie, S.: Monaural speech separation and recognition challenge, *Computer Speech and Language*, vol. 24, pp. 1–15 (2010).
- [19] Vincent, E., Araki, S., Theis, F.J., Nolte, G., Bofill, P., Sawada, H., Ozerov, A., Gowreesunker, B.V., Lutter, D., and Duong, N.Q.K.: The Signal Separation Evaluation Campaign (2007-2010): Achievements and remaining challenges, *Signal Processing*, vol. 92, pp. 1928–1936 (2012).
- [20] Delcroix, M., Kinoshita, K., Nakatani, T., Araki, S., Ogawa, A., Hori, T., Watanabe, S., Fujimoto, M., Yoshioka, T., Oba, T., Kubo, Y., Souden, M., Hahm, S.J., and Nakamura, A.: Speech recognition in living rooms: Integrated speech enhancement and recognition system based on spatial, spectral & temporal modeling of sounds, *Computer Speech and Language*, Elsevier, vol. 27, no. 3, pp. 851–873 (2013).
- [21] 堀貴明, 小川厚徳, 藤本雅清, 大庭隆伸, 久保陽太郎, ハム ソンジュン, 荒木章子, ソウデン メレズ, デルクロア マーク, 吉岡拓也, 木下慶介, 中谷智広, 中村篤: 会話分析タスクにおける複数人自由会話音声認識の改善, 日本音響学会秋季研究発表会, pp. 55–56, 9 月 (2012).
- [22] Kinoshita, K., Delcroix, M., Souden, M., and Nakatani, T.: Example-based speech enhancement with joint utilization of spatial, spectral & temporal cues of speech and noise, *Proc. Interspeech* (2012).
- [23] Ishiguro, K., Yamada, T., Araki, S., Nakatani, T., and Sawada, H.: Probabilistic speaker diarization with bag-of-words representations of speaker angle information, *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 20, no. 2, pp. 447–460 (2012).