〈研究発表〉

高雑音下での声帯マイクを用いた音声認識

天野 亘[1)]，野口 健[1)]，武田 龍[2)]，本間 健[2)]

[1)] 東京都下水道サービス㈱ （〒100-0004 東京都千代田区大手町2-6-2
E-mail : wataru-amano@tgs-sw.co.jp, ken-noguchi@tgs-sw.co.jp）

[2)] ㈱日立製作所 中央研究所 （〒185-8601 東京都国分寺市東恋ヶ窪1-280
E-mail : ryu.takeda.qh@hitachi.com, takeshi.homma.ps@hitachi.com）

概 要

本稿では，高雑音下でも正確な入力を可能にする声帯マイクを用いた音声認識システムの開発に関して報告を行う。声帯マイクは咽喉の振動を直接取得するので，騒音環境下でも利用者の声だけをシステムに入力できる。しかし，声帯マイク音声は気導マイク音声と音響的特徴が異なるため，通常の音声認識システムでは認識が難しい。本研究では，声帯マイク用音響モデルの構築，モデル適応技術を用いた解決を行う。下水処理施設の設備運転時における高雑音下で収録した音声により，93% の認識率を達成し，本システムの有効性を確認した。

キーワード：音声認識，声帯マイク，モデル適応，設備点検，騒音環境

1. はじめに

下水処理場では，設備を維持管理するための点検が日々行われている。この点検では，複数の点検員によって，設備の整備・清掃のほか，計器の確認と点検表への記録が行われる。点検の手順や順路，点検のポイントなど，点検員が長年培った知識や経験によって，効果的な点検が効率的になされている。一方，技術者の減少や高齢化により，知識・経験の若手への継承や作業の質の維持が，課題となっている。これらの課題解決に向け，ICT 技術を活用した点検サポートツールの開発が求められていた。

本研究では，タブレットなどの携帯端末上に音声認識システムを構築し，点検表への音声入力の実現を目指すものである。これまでの点検表への手書き入力とは異なり，音声入力を採用することで，点検員が発話するだけで点検表が完成するもので，点検のハンズフリー化により安全性の向上と点検のレベルアップを図るとともに，点検データが即座に電子データ化されるので，設備の計画的な更新や維持管理への活用が期待できる（**Fig. 1**）。

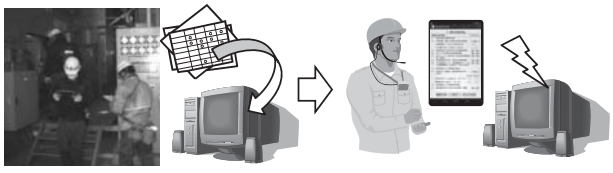

Fig. 1 研究開発の目的

2. 高雑音下での音声認識

音声認識システムを下水処理場で用いるためには，1)高雑音下で音声認識が可能，2)携帯端末上で動作可能，という要件を満たす必要がある。これは，下水処理場には，ガスタービン設備といった騒音レベルが 90dBA を超す設備があるためである。また，地下など遮蔽された空間に設備があることが多く，通信環境も整備されていないため，携帯端末と外部サーバ間との通信は困難である。そのため，音声認識を含め，すべての処理はリソースが制限された携帯端末上で動作できることが望ましい。

雑音下を前提とした音声認識システムでは，一般的に，マイク入力信号から雑音信号を抑圧してから音声認識を行う。雑音を抑圧する方法には，複数のマイク利用を前提とした方法（アレイ処理）[1,2)]と，1つのマイクを前提とした方法に大別できる[3,4)]。前者は，雑音や残響に指向性を適応することで，高品質な音声信号が得られるが，1)アレイ実装（複数マイクの実装）が必要，2)演算量が膨大，という点で，携帯端末には適さない。後者は，演算量と雑音抑圧性能のバランスが良く，携帯端末への実装に適している。しかし，両者に共通する問題として，雑音抑圧結果が信号対雑音比（SN 比）に大きく依存する点がある。そのため，SN 比が極端に悪い環境やユーザの声量が小さい場合では，雑音抑圧性能の劣化を避けられない。

本研究では，音声入力用のマイクに声帯マイク（**Fig. 2**）を用いる方法に着目する。声帯マイクは，咽喉の振動を直接取得する。そのため，服との摩擦音と

Fig. 2　声帯マイク

いった雑音を除けば，ガスタービン設備といった騒音環境下でも，利用者の音声をより選択的に入力できる。しかし，音声認識システムは通常，気導マイクで収録した音声から学習により構築した音響モデルに基づき，認識処理を行う。気導マイクと声帯マイクでは，音声信号の周波数特性に違いがでるため，気導マイク用の音響モデルで声帯マイクの音声を認識すると，モデルの不一致により認識率が低下する問題がある。

今回，モデル不一致の問題に対し，声帯マイク用の音響モデルを，適応技術を用いて構築することで解決を図った。声帯マイク用音響モデルを構築するには，大量の声帯マイク音声データのみを用いて構築することが理想であるが，様々な話者・文章を含む 100 時間以上の音声データを収集し，一から音響モデルを構築するのは高コストである（**Fig. 6a**）。一方，気導マイク用の音響モデルは一般に販売されており，話者・文章の種類，分量ともに豊富である。モデル適応技術を用いれば，少量の声帯マイク音声データと気導マイク用音響モデルから，効率的に声帯マイク用音響モデルを構築できる。これらの方法を用いて構築した音声入力システムを，実際の下水道設備の点検環境で収録した音声での数値認識率評価を行うこととした。

声帯マイクと気導マイクのペアを用いたアプローチはいくつか報告されている[5,6]。しかし，本研究のように，不特定話者を想定し，ある程度の規模のデータベースを用いたモデル適応によるアプローチは，これまであまり見られない手法である。

3. 音声認識の原理とシステム設計

3.1　音声認識の概要

音声認識の概要を説明する（**Fig. 3**）。音声認識を実現するためには，事前に，複数の話者による多数の学習用音声データ，時系列の音声特徴量抽出，音響モデルの構築，言語モデルの構築といった『学習フェーズ』が必要である。

時系列の音声特徴量とは，短時間周波数成分を数値化したものであり，また，音響モデルとは，音声特徴量と音の表記を対応づけたデータベースである。具体的には，「あ」の音の特徴量はこのような分布，「い」の音は……，という統計的モデルである。この対応付けは，機械学習によって多数の音声と読み仮名を用いて自動的に構築される。言語モデルは文法に近いものであり，単語辞書と単語と単語の繋がりルールを記録したデータベースである。

人の音声を認識する『認識フェーズ』では，まず，入力音声の音声特徴量を抽出する。次に，言語モデルから次に発話される単語が予測される。さらに，単語の読みと音響モデルから，その候補が入力音声にどれだけ近いかが評価される。これにより，入力された音声に対してもっともらしい単語列が出力される。

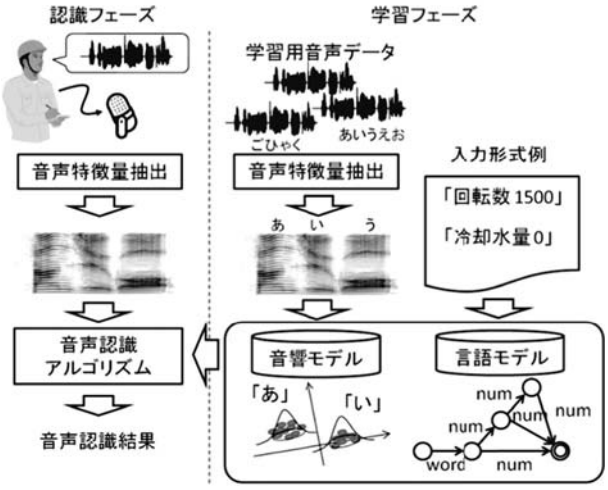

Fig. 3　音声認識の概要

3.2　点検向け音声入力システムの基本設計

今回試作した音声入力システムは，実際の下水処理場で使われている点検表をベースにしており，紙媒体の形式をそのままタブレット画面に表示する。また，ボタンを押すことで，認識の開始と終了を行う。

入力データは，「項目名（キーワード）」と「数値」，および，「良否」項目が主である。例えば，「機関入口温度 23」「回転数 1500」などである。このような項目を認識できるように，単語辞書と繋がりのルールをグラフで表現し，計器の数値範囲を配慮した言語モデルを構築した。

3.3　下水道施設内騒音環境下と声帯マイクの課題

音声認識を行うに当たり，現場の下水道施設内における騒音レベルを調査した。設備が稼働していない場所に関しては，約 70dBA の騒音レベルであり，比較的静かであった。一方，設備が稼動している場所では，80dBA 以上の騒音レベルを示す場所もあった。特に，ガスタービン設備，ポンプ設備（**Fig. 4**）の点検環境では，回転等による動作音により，高い騒音レベルを

示していた。このような騒音下での音声認識が必要となる。

騒音環境下で音声認識を達成するため，本研究では，周囲の騒音が入力されにくい声帯マイクを利用する。声帯マイクには，外部の騒音が混入しにくいメリットがある。一方，声帯マイクと気導マイクで収録した音声は音響的特徴が異なっている（**Fig. 5**）ため，声帯マイクの音声特徴に適合した音響モデルを構築することが課題となる。

Fig. 4 ガスタービン設備（左）とポンプ設備（右）

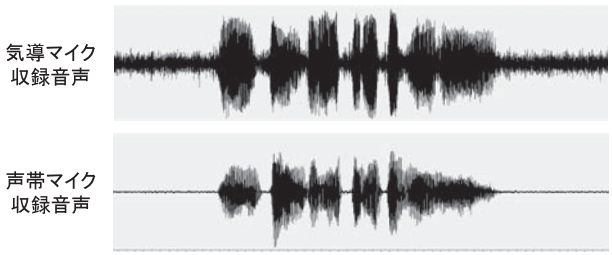

Fig. 5 気導マイク音声（上）と声帯マイク音声（下）の波形

4. 声帯マイク用の音響モデルの構築

声帯マイク用の音響モデルの構築にあたって，本研究では，数時間規模の声帯マイク音声を用いて，気導マイクで構築した音響モデルを声帯マイク用に適応し，再利用するアプローチを取る（**Fig. 6c**）。

そのために，まず，男女含む不特定話者 24 名より，約 15 時間規模の声帯マイク音声を収集し，音響モデル適応用のデータとした。音響モデルを適応する方法として，最大事後確率法を採用する。最大事後確率法は，適応用データを最もよく表現するようにモデルパラメータ（平均・分散など）を更新する方法であり，適応用データの量に応じて更新前のモデルパラメータの寄与度を調節する。例えば，適応用データが少ない場合は，適応前のモデルパラメータから大きく変更されないようにし，データ量が多い場合は適応用データに適合するようにパラメータが更新される（**Fig. 7**）。今回は，複数の多次元ガウス分布の 1 つから音声特徴量が生成されることを前提とした Gaussian Mixture Model（GMM）を採用した[8]。

Fig. 8 に気導マイクと声帯マイクの音声の時間–周波数成分（スペクトログラム）を示す。この音声は，同一の発話を 2 つのマイクで同時に収録したデータである。これをみると，両者で同じような周波数成分を含む部分もあれば，両者の周波数成分が異なる部分も存在している。特に，「ん」の発音，及び，子音部分に特徴の違いが大きくみられた。そのため，似た特徴部分に関しては，気導マイク音響モデルのパラメータを再利用し，異なる特徴部分に関しては声帯マイク音声の特徴をよく表現するようにパラメータ（平均と分散）が変化することが期待できる。

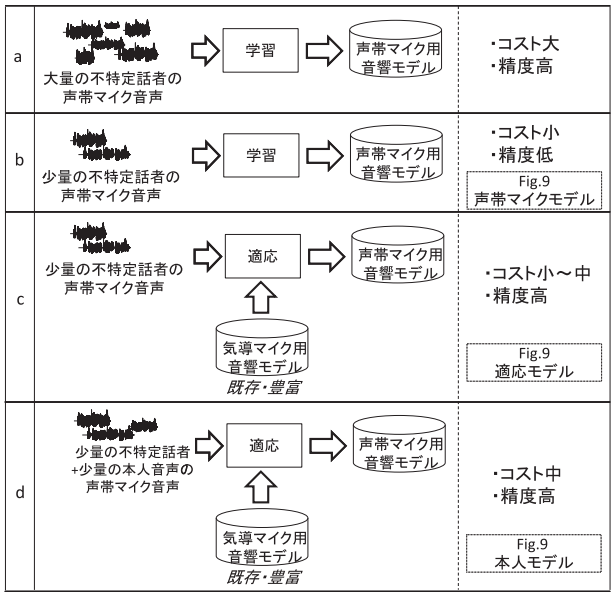

Fig. 6 声帯マイク用の音響モデル構築パターン

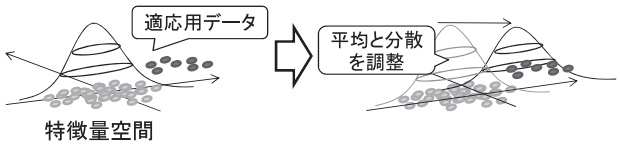

Fig. 7 特徴量分布とモデルパラメータ適応

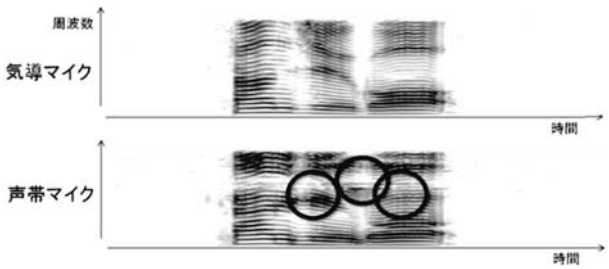

Fig. 8 気導マイク音声（上）と声帯マイク音声（下）のスペクトルグラム（横軸は時間，縦軸は周波数）

5. 評　価　実　験

5.1　実験設定

下水処理施設で収録した声帯マイクの数値発話の音声を用いて，その認識率を評価した。施設内の 5 つの環境下（ガスタービン設備含む）で音声を収録した。収録環境の騒音レベルはそれぞれ 89, 89, 85, 80, 70 dBA であった。特徴量は Mel-Frequency Cepstrum Coefficience（MFCC）とその 1 次と 2 次差分を加え，ケプストラム平均除去[9]を施したものを採用した。

音響モデルは一般的に用いられる状態共有型・3 状態トライフォン GMM-HMM を用いた。評価用セットは，桁数がそれぞれ 1, 2, 3, 4 の棒読みと桁読みを含む数値を約 100 個用意し，1 個の発話に対して 2 回ずつ収録した。サンプリング周波数は，8kHz とした。認識された数値桁がすべて正解と合致したときのみ，正解として扱い，正解率を算出した。

まず，声帯マイク音声に対し，約 100 時間の不特定話者音声で学習した既存の気導マイクモデル，15 時間データから構築した声帯マイクモデル（**Fig. 6b**），および，気導マイクモデルを声帯マイクに適応させたモデル（**Fig. 6c**）による認識結果を比較する。さらに，気導マイク音声と 1 マイク雑音抑圧処理[4]を行い，気導マイクモデルで認識した結果も比較する。点検者本人の音声を事前に収録できることを想定し，不特定話者データ 40 分と評価者のデータを 20 分程度用いて適応した本人モデル（**Fig. 6d**）結果も併せて示す。

評価対象の発話者は，不特定話者における学習データでは音声収録していない男性 2 名である。

5.2　実験結果および考察

Fig. 9 に音声認識結果を示す。まず，気導マイクと気導マイクモデルを用いて，雑音抑圧をしない場合の正解率は 60〜80% 程度であった。これに対し，気導マイク音声と雑音抑圧の組み合わせでは，雑音抑圧が機能しており，話者 A に関する正解率は高い（94.2%）。一方，話者 B では，正解率が改善しているものの，雑音レベルの影響や近隣にいる別の点検者の音声信号の影響により，正解率が十分に向上していない（85.7%）。そのため，雑音抑圧処理を加えるだけでは，騒音レベルに対して安定した正解率を達成することは難しいといえる。

入力	音響モデル	平均正解率（%）	環境騒音レベル（dBA）				
			89	*89*	*85*	*80*	*70*
気導マイク	気導マイクモデル	82.3	80.1	71.4	70.5	93.6	92.5
気導マイク＋雑音抑圧	気導マイクモデル	94.2	92.5	93.1	92.0	96.6	97.1
声帯マイク	気導マイクモデル	66.2	67.2	71.4	61.5	65.6	65.6
	声帯マイクモデル	91.1	89.1	90.6	89.1	90.1	96.9
	適応モデル	93.5	91.7	92.2	93.8	93.2	96.9
	本人モデル	98.9	97.9	99.0	100.0	99.0	98.4

入力	音響モデル	平均正解率（%）	環境騒音レベル（dBA）				
			89	*89*	*85*	*80*	*70*
気導マイク	気導マイクモデル	62.1	37.0	46.0	57.4	73.8	88.5
気導マイク＋雑音抑圧	気導マイクモデル	85.7	82.1	76.4	87.2	89.5	93.5
声帯マイク	気導マイクモデル	78.4	83.9	82.3	76.0	74.0	76.0
	声帯マイクモデル	89.3	86.5	86.5	88.0	88.5	96.9
	適応モデル	93.8	94.8	90.6	93.8	93.8	95.8
	本人モデル	97.2	97.7	94.3	98.9	99.4	96.0

Fig. 9　話者 A（上）と話者 B（下）の正解率（%）

次に，声帯マイク音声の結果を比較する。気導マイクモデルでは正解率は，話者 A が 66.2%，話者 B が 78.4% であった。声帯マイクモデルでは，正解率はより向上し，話者 A が 91.1%，話者 B が 89.3% であった。気導マイクモデルに声帯マイク音声で適応を行う条件にて，もっとも高い正解率が得られ，話者 A で 93.5%，話者 B で 93.8% を達成できた。声帯マイクモデルより正解率が向上した原因としては，声帯マイクモデルだけでは十分なデータ量がなく，正解率の向上への寄与が限定されていたためだと考えられる。

最後に，本人音声をモデル適応に用いた場合は，さらに高い正解率が出ている（話者 A 98.9%，話者 B 97.2%）。これは，実運用を想定した場合，運用前に，実際の点検者による少量の音声データを収集するという，わずかな手間で，運用時の認識率を向上させ，使い勝手の向上に大きく寄与するものである。

以上の結果より，声帯マイクと音響モデル適応を用いた音声認識により，高騒音環境下で実用に耐えうる音声認識が可能であることがわかった。

6. お　わ　り　に

本稿では，下水処理場設備向け点検サポートツールの一環として，声帯マイクを用いた点検表の音声入力について報告した。気導マイクと声帯マイクの音声特徴量の違いによる認識率低下の問題を適応技術により解決した。また，実際の環境で収録した音声でその有効性が確認できた。この技術は，下水道施設はもちろん，強い環境騒音下での音声認識技術として幅広く応用の利く技術である。今後は構築したタブレット端末の現場で利便性向上のために，インタフェース等の改良を行うことで，実際の点検作業に適用する予定である。

参 考 文 献

1）　L. J. Griffiths and C. W. Jim : “An alternative approach to linearly constrained adaptive beamforming,” *IEEE Trans. Antennas and Propagation*, Vol. 30, No. 1, pp. 27–34, 1982.
2）　M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, N. Nukaga :

"Optimized Speech Dereverberation From Probabilistic Perspective for Time Varying Acoustic Transfer Function," *IEEE Trans. on ASLP*, Vol. 21(7), pp. 1369–1380, 2013.
3) I. Cohen and B. Berdugo : "Speech enhancement for Non-stationary Noise Environments," *Signal Processing*, Vol. 81, pp. 2403–2418, 2001.
4) Y. Obuchi, R. Takeda and M. Togami : "Bidirectional OM-LSA Speech Estimator for Noise Robust Speech Recognition," *proc of ASRU*, pp. 173–178, 2011.
5) P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari and K. Shikano : "Accurate hidden Markov models for non-audible murmur (NAM) recognition based on iterative supervised adaptation," *proc of ASRU*, pp. 73–76, 2003.
6) E. Erzin : "Improving Throat Microphone Speech Recognition by Joint Analysis of Throat and Acoustic Microphone Recordings," *IEEE Trans. on ASLP*, Vol. 17, No. 7, pp. 1316–1324, 2009.
7) 安藤彰男：リアルタイム音声認識，電子情報通信学会，2003.
8) J.-L. Gauvain, and C.-H. Lee : "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, Vol. 2, No. 2, pp. 291–298, 1994.
9) S. Furui : "Cepstral analysis technique for automatic speaker verication," *IEEE Trans. on ASLP*, Vol. 29, pp. 254–272, 1981.