様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　２４　日現在

機関番号：　３３９３４
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2017
課題番号：　２６３３０２１１

研究課題名（和文）音声対話システムを対象とした雑音に頑健な話者までの距離推定の研究

研究課題名（英文）Robust Method of Distance Estimation to a Speaker for Spoken Dialog System

研究代表者
實廣　貴敏（Jitsuhiro, Takatoshi）

愛知工科大学・工学部・准教授（移行）

研究者番号：６０３９４９９６

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：音声対話システムにおいて，周囲状況を把握する方法の一つとして，単一マイクロホンにおいて，音声そのものからその特性を推定・識別することで，発話者の口からマイクまでのおよその距離を推定する手法を提案する．距離ごとに収録された音声データをDeep Neural Network (DNN)の一種で学習する．使用時には，短時間に区切られた音声フレームをDNNに入力し，推定距離を出力する．全フレームで推定距離の多数決を行うことで1発話の推定距離を得る．0.2 mと5 mの音声識別実験では，約85%の識別率を得ることができた．

研究成果の概要（英文）：We propose the estimation method of distance from a mouth of a speaker to a microphone by estimating and classifying the feature of speech recorded by a single microphone. A Deep Neural Network (DNN) is training using speech data recorded for each distance. For estimation, short-time speech frames are entered into the DNN, it will estimate the distance for each frame. After that, the estimated distance is obtained for one utterance by majority decision of estimated distance in all frames. In speech recognition experiments of 1 m and 5 m, the proposed method can obtain about 85 % identification rate.

研究分野：　音声情報処理

キーワード：　音声認識　音声対話システム　音源距離推定　音響モデル　VQコードブック　深層学習　Deep Belief Network

１．研究開始当初の背景

音声認識技術による音声対話システムでは，現状，入力される音声が音声認識対象として認識される．しかし，実際に使う場所では，周囲の雑音や認識対象ではない音声が観測される．一般に，入力パワーが，ある閾値より大きく，音声と雑音を識別するモデルにより音声のみを抽出し，音声として認識対象とする．だが，それ以上の考慮はあまりされていない．音声対話システムを実環境で使えるようにするには，周囲の環境をより正確に把握し，対応できるようにする必要がある．

本課題では，周囲状況推定技術の一つとして，マイクロホンから発話者までのおよその距離推定技術の確立を主な目的とする．発話者の位置がマイクロホンの近距離なら音声認識対象である可能性が高く，遠距離なら音声認識対象外である可能性が高いと考えられる．

音源位置を複数マイクロホンで推定する技術は，三角測量によるもの[1] [2]，小型正十二面体マイクロホンアレイを用いたもの[3]などがある．しかし，音声対話システムにおいては単一マイクロホンが使われることが多い．そこで，研究代表者らは[4]の中で，対象は音声に限定されるが，単一マイクロホンからの入力音声を用い，話者までの距離を推定する手法を提案した．ただし，この検討では比較的雑音が少ない状態で，また，音響空間特性はインパルス応答を評価データに畳み込んで作成したシミュレーションによるものであった．

２．研究の目的

音声対話システムを実環境で使えるようにするには，周囲の環境をより正確に把握し，対応できるようにする必要がある．本課題では，周囲状況推定技術の一つとして，単一のマイクロホンから発話者までのおよその距離推定技術の確立を主な目的とする．発話者の位置がマイクロホンの近距離なら音声認識対象である可能性が高く，遠距離なら音声認識対象外である可能性がある．本研究では，実際の環境で収録した音声データを用い，その環境下での提案方法[李ら，音講論, 2013]の評価を行う．また，近年のパターン認識技術で発展してきた Deep Neural Network (DNN) 技術を本研究課題に適用し，よりロバストな推定技術の確立を目指す．

３．研究の方法

本課題では，研究の進度により，大きく 2 種類の手法で距離推定を検討した．一つ目は[4]で提案している(1) VQ コードブックを用いた手法で，実環境音声データにおいて評価した．二つ目は DNN の一つである(2) Deep Belief Network (DBN) [5]を用いた手法である．まず，それぞれの手法をまとめておく．

(1) VQ コードブックを用いた距離推定手法

あらかじめ，各距離からマイクロホンまでの音響伝達特性をテンプレート化しておく．入力音声から推定した音響伝達特性とテンプレートを比較し，最も近いテンプレートの位置が，推定された発話者の位置となる．

次に，入力音声から音響伝達特性を推定する方法について述べる．時刻tにおける入力音声のスペクトル$\boldsymbol{X}_t$は下記のように表される．

$$\boldsymbol{X}_t = \boldsymbol{H}_t \cdot \boldsymbol{S}_t + \boldsymbol{N}_t \quad (1)$$

ここで，$\boldsymbol{S}_t, \boldsymbol{H}_t, \boldsymbol{N}_t$はそれぞれクリーン音声のスペクトル，音響伝達周波数特性，加算性雑音のスペクトルである．加算性雑音$\boldsymbol{N}_t$は無視できるほど抑圧できているとすると，

$$\boldsymbol{X}_t \approx \boldsymbol{H}_t \cdot \boldsymbol{S}_t \quad (2)$$

と近似できる．両辺の対数を取ると推定される対数音響伝達特性は，

$$\log \widehat{\boldsymbol{H}}_t \approx \log \boldsymbol{X}_t - \log \boldsymbol{S}_t \quad (3)$$

と表せる．$\log \boldsymbol{S}_t$が推定できれば，$\log \widehat{\boldsymbol{H}}_t$を求めることができる．$\log \boldsymbol{S}_t$を推定するために，クリーン音声データベースから VQ コードブックを作成しておき，入力に最も近いセントロイド・ベクトル$\log \boldsymbol{C}_i$を探し，それを$\log \boldsymbol{S}_t$とする．ここでは距離を下記のように定義する．

$$d_i = (\log \boldsymbol{X}_t - \log \boldsymbol{C}_i)'(\log \boldsymbol{X}_t - \log \boldsymbol{C}_i) \quad (4)$$

‘’は転置を示す．この値が最も小さくなる$\log \boldsymbol{C}_i$を推定されたクリーン音声$\log \widehat{\boldsymbol{S}}_t$とする．

$$\log \widehat{\boldsymbol{H}}_t \approx \log \boldsymbol{X}_t - \log \widehat{\boldsymbol{S}}_t \quad (5)$$

$\log \widehat{\boldsymbol{S}}_t$は元の音声の特性と大まかなところでは一致すると考えられ，入力音声との差分がその空間での音響伝達特性となる．ただし，1 フレームごとの推定では不安定になると考えられる．そこで，対数音響伝達特性$\log \bar{\boldsymbol{H}}$は数秒程度の 1 発話においては一定と仮定できるので，全てのフレームに対する平均として推定できる．

$$\log \bar{\boldsymbol{H}} = \frac{1}{T} \sum_{t=1}^{T} \log \widehat{\boldsymbol{H}}_t \quad (6)$$

ここで，Tは総フレーム数である．このアルゴリズムを音声区間に適用することにより，安定に音響伝達特性を推定できる．さらに，一度得た$\bar{\boldsymbol{H}}$からクリーン音声を推定し，同じアルゴリズムで再度$\bar{\boldsymbol{H}}$を推定，それを何度か繰り返す繰り返し推定を用いる．

(2) Deep Belief Network (DBN)を用いた手法

近年，パターン認識をはじめ，多くの分野で特によく利用されるようになった Deep Neural Network (DNN)がある．本研究も一つのパターン認識タスクではあるので，これらの技術を適用する．今回は基本的な技術の一つ, Deep Belief Network (DBN)を用いる．

音声はその継続時間長は様々なため，一度にニューラルネットワークに入力して扱うことが難しい．そのため，短時間フレームに分け，それらを 1 フレームずつ DNN で認識処理を行い，さらに別処理を行うことが多い．本研究では，各フレームごとに DBN を使って距離推定を行い，(a) 各フレームでの距離別スコアを累積し，音声終端で累積スコアの最も大きい距離を選択する手法，(b) 各フレームで推定された距離を音声終端で多数決により

1 発声に対する推定距離を決める手法，の 2 つについて検討した．

４．研究成果

実際には「５．主な発表論文等 〔その他〕」にあるように，主に卒業研究の一環として，多くの検討を行った．その全てをこの資料にまとめることは困難であるので，代表的な結果のみをまとめておく．

(1) 音声収録環境

引用文献[4]およびその時点での検討を詳細にまとめた〔雑誌論文〕①では，対象とする場所にてインパルス応答のみを収録し，それをコンピュータ上での畳み込みにより得られたシミュレート音声に対する評価であった．今回は，実環境において音声を再生録音したものに対して評価を行った．

図 1 に収録環境を示す．床から天井までの高さが 275 cm である大学内の教室にて，音声データをスピーカから再生し，マイクロホンで収録した．マイクロホンは固定し，スピーカのマイクロホンからの位置を 0.2，0.5，1，2，3，4，5，6 m と変えて録音を行った．マイクロホンは Sony 製 C-357，スピーカは YAMAHA 製 MSP7STUDIO，収録機器は Roland 製 R-44 を用いた．マイクロホンとスピーカの中心の高さは 1.3 m とした．収録でのサンプリング周波数は 48 kHz，量子化ビット数を 16 bit とした．実験で使用する際には，16 kHz にダウンサンプリングをして用いた．

再生録音した音声データとして，日本音響学会新聞記事読み上げ音声コーパス(Japanese Newspaper Article Sentences: JNAS)から，男女各 100 発話(男性 23 人，女性 23 人)を用いた．

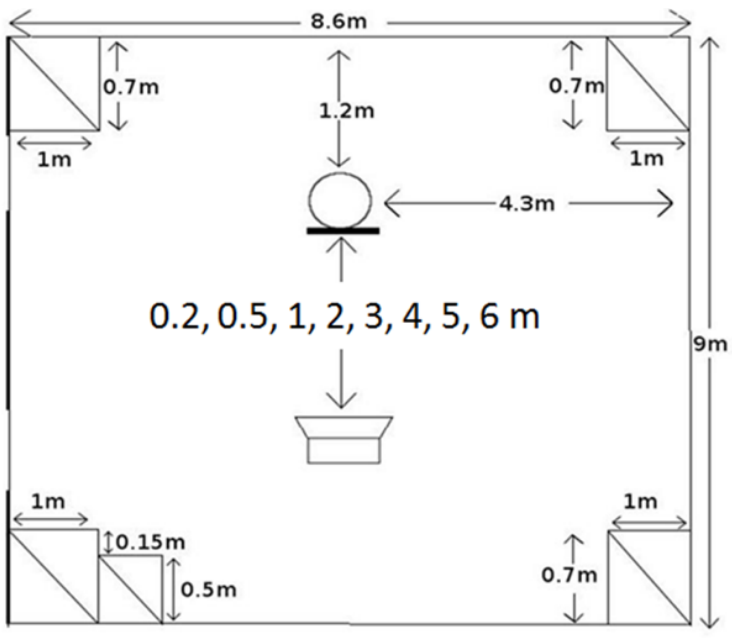

図 1 収録環境

(2) 手法(1)：VQ コードブックを用いた距離推定手法での成果

クリーンな音声を表す VQ コードブックを音声コーパス JNAS のクリーン音声を用いて k-means クラスタリングにより生成した．音声特徴量は音声認識でよく利用されて来た MFCC 13 次元のみとした．いくつか条件を変えて検討を行ったが，ここでの結果は，男性 14 人，女性 15 人を含む計 1,100 発話を学習データとして用い，コードブックサイズを 256, 512, 1024, 2048, 4096 として比較を行った場合のものである．距離推定認識率を図 2 に示す．距離が 0.2 m では平均 94%，0.5 m では平均 2.4%が得られた．それ以外の距離では認識率は 0%に近い値となった．最も近い距離では推定がうまく働くが，それ以外では，正しく識別できないようである．特徴量もより多くの種類で検証していく必要がある．

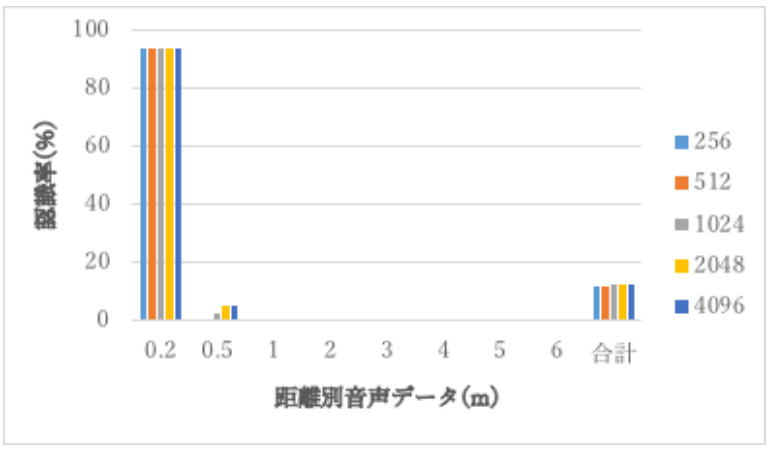

図 2 VQ コードブックを用いた距離推定手法による認識結果

(3) 手法(2)：Deep Belief Network (DBN)を用いた手法

VQ コードブックを用いた手法では，改善が困難と考えられるので，大きく手法自体を検討し直すこととし，DNN の技術を利用した．今回は基本的なニューラルネットワークである DBN を用いた．音声データは図 1 で収録されたものを用い，男性 23 話者 100 発話の内，男性 11 話者 50 発話を評価データ，男性 9 話者 38 発話を学習データ，男性 3 話者 12 発話を検証データに，また，女性 23 話者 100 発話の内，女性 13 話者 52 発話を評価データ，女性 7 話者 35 発話を学習データ，女性 3 話者 13 発話を検証データに用いた．音声特徴量として 1 フレームあたり 12 次元 MFCC，12 次元 Δ MFCC，Δ 対数パワーを用いた．それらを前後 11 フレーム分を接続し，275 次元を DBN への入力とした．今回の距離推定では，基本的な性能を見るため，0.2 m と 5 m の 2 種類のみとした．ただし，DBN への入力は無音区間も存在するため，出力は無音，0.2 m，5 m の 3 種類とした．隠れ層の階層数やそれぞれのユニット数は様々にふって適切なパラメータを見つけることにする．DBN は 1 フレームごとに識別することになるが，それらを(a) 各フレームでの距離別対数スコアを累積し，音声終端で最も大きい累積対数スコアの距離を選択する場合，(b) 音声終端で多数決により推定距離を決める場合，の検討を行った．図 3 に(a)手法，図 4 に(b)手法での結果を示す．各マーカーが一つのモデルによる評価値である．適合率と再現率による評価で，右上にあるほど精度が高いことを示している．

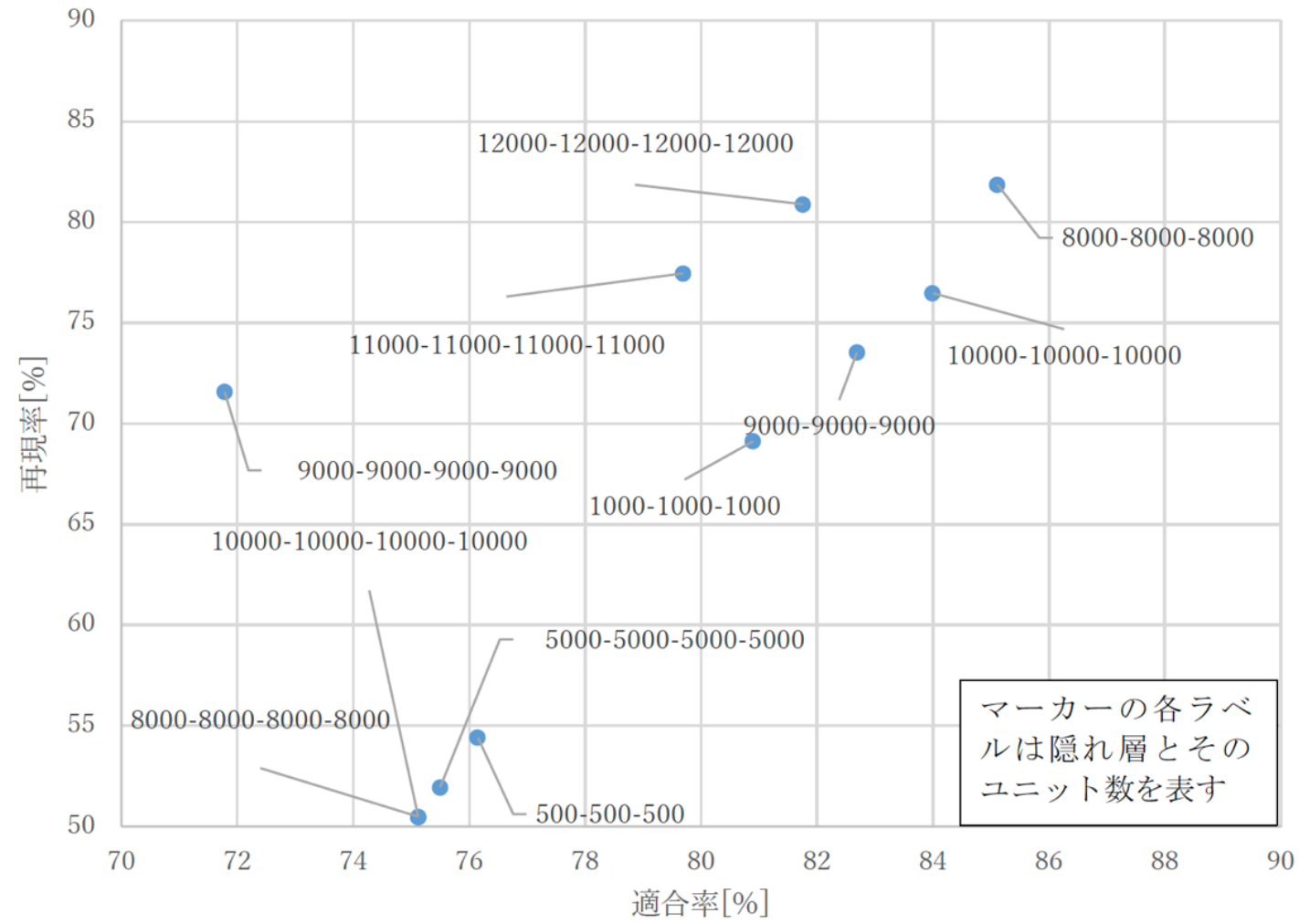

図 3 (a) 累積対数スコアによる 2 距離推定での適合率と再現率

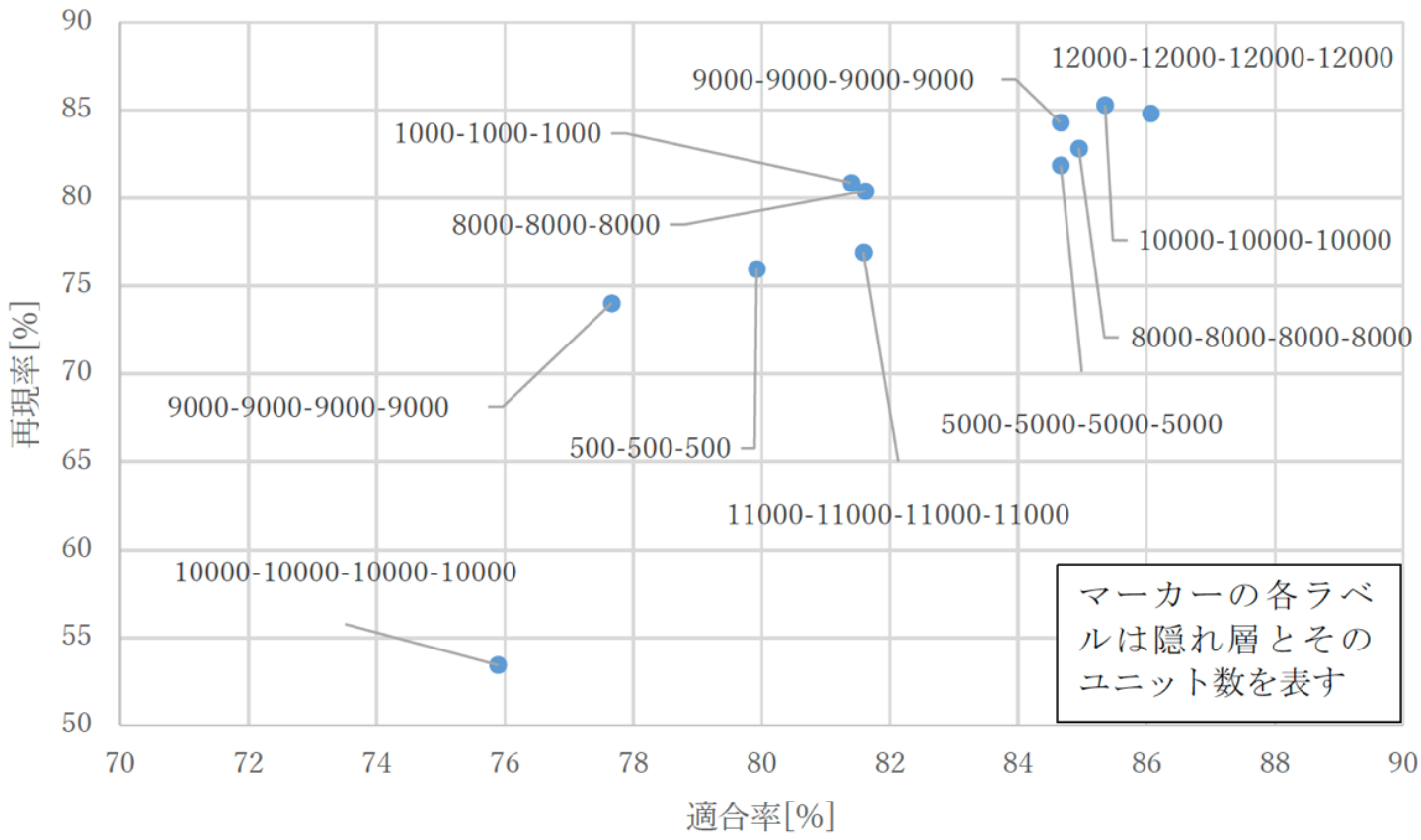

図 4 (b) 多数決による 2 距離推定での適合率と再現率

マーカーに付いている数値列は，数値の数が隠れ層数，各数値が各層のユニット数を表している．例えば，「500-500-500」であれば，隠れ層数は 3，各層のユニット数は 500，であることを示す．ユニット数を 1 万ユニットと，この手のニューラルネットワークとしては，ユニット数が大きくないと高い精度が得られないことがわかった．(a) 累積対数スコアによる手法では，隠れ層数 3(全体では 5 階層)，各層 8000 ユニットで最もよく，適合率 85.11%，再現率 81.86%，F 値 83.46%となった．(b) 多数決による手法では，隠れ層数 4 (全体では 6 階層)，各層 12000 ユニットで最もよく，適合率 86.06%，再現率 84.81%，F 値 85.43%となった．全体的に(b) 多数決による手法が良い精度を得られやすいことがわかった．

また，最適なパラメータは特に傾向が見られず，階層数とユニット数の組み合わせをできるだけ多く試す必要があることがわかった．

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

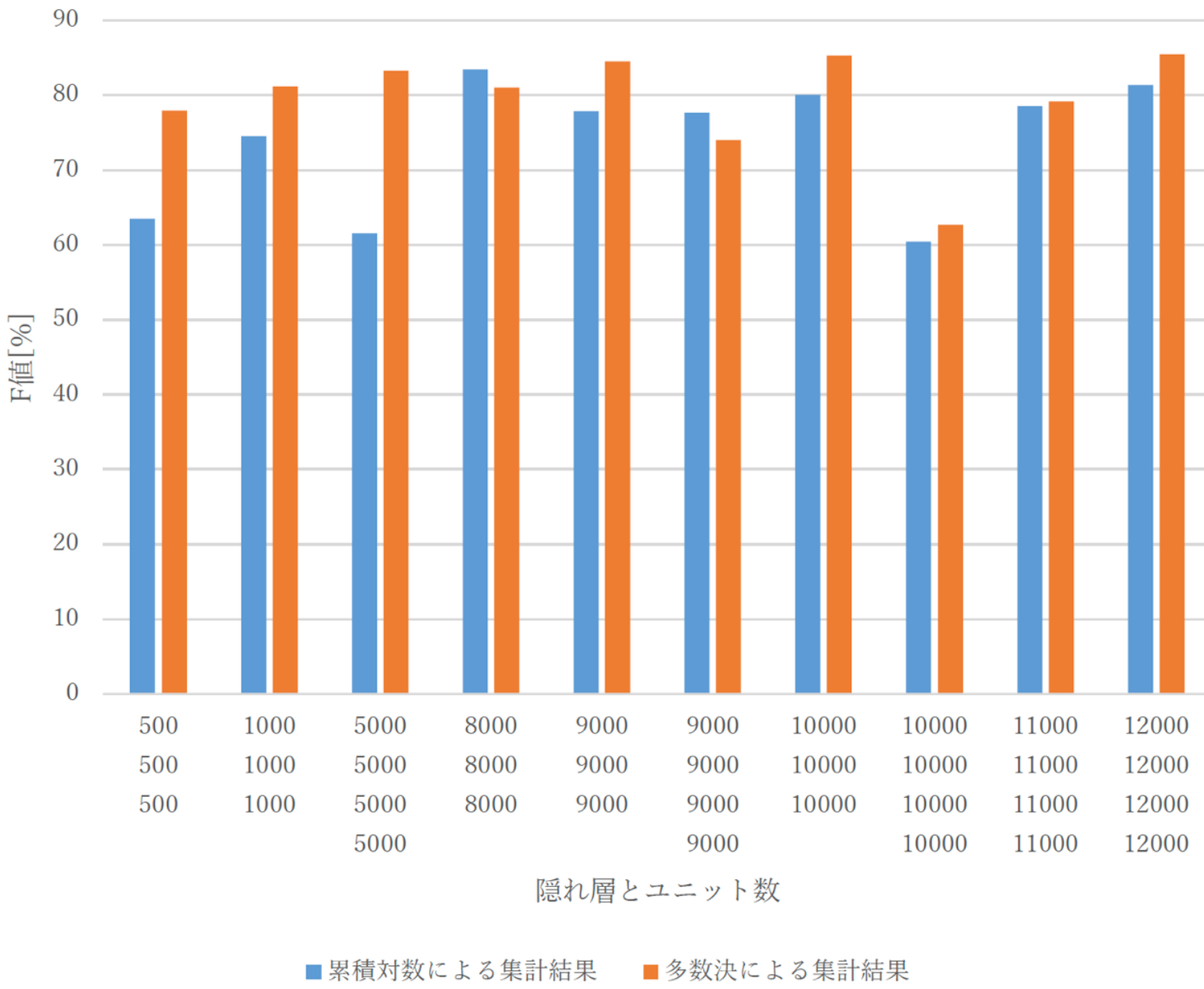

図 5　2 つの集計方法による F 値の比較

わかりやすく比較するために，この 2 つの集計方法による同じモデルパラメータでの F 値を図 5 に示す．同じモデルパラメータでは，隠れ層数 3，各層ユニット数 8000 の時，または，隠れ層数 4，各層ユニット数 9000 の時にのみ，累積対数スコアによる F 値が多数決による手法に比べ，精度が上回った．それ以外では，多数決による手法が上回った．

得られた結果の詳細を分析すると，無音区間で，5 m と誤推定することが多く，累積対数スコアでは，無音区間で累積されるスコアが多くなり，全体として誤推定につながると考えられる．

ユニット数が大変多くなることから，特徴量の工夫がまだまだ必要と考えられる．今回，MFCC ベースの特徴量を用いたが，より直接的に音声の特徴を含んでいる周波数スペクトルそのものを使うことが可能と考えられる．また，今回は 11 フレームの連結を行なったが，マイクから離れるほど，反射波の影響が大きくなるため，より長い範囲を考慮できる枠組みを考える必要がある．さらに，時間方法の変動を考慮できる RNN 系のニューラルネットワークを利用することも考えられる．

また，〔学会発表〕②は，本研究が着目している音声対話システムでの距離推定を音声ではなく，Bluetooth を利用して行うアイディアを実現したものである．

＜引用文献＞
[1] M. S. Brandstein, H. F. Silverman, “A practical methodology for speech source localization with microphone arrays,” Computer Speech & Language, Vol. 11, No.2,pp.91-126,1997.
[2] P. Bergamo, S. Asgari, H. Wang, D. Maniezzo, L. Yip, R. E. Hudson, K. yao, D. Estrin, “Collaborative sensor networking towards real-time acoustical beam-forming in free-space and limited reverberance,”IEEE Trans. Mobile Computing, Vol. 3, Issue 3, 2004.
[3] Satoshi Esaki, Kenta Niwa, Takanori Nishino, Kazuya Takeda, “Estimating sound source depth using a small-size array,” Proceedings of ICASSP 2012, pp401-404, 2012.
[4] 李津，實廣貴敏，武田一哉，”単一マイクロホンによる音響モデルを用いた発話者までの距離推定,” 日本音響学会 2013 年春季研究発表会講演論文集, pp. 29-30, 2013 年
[5] G.E.Hinton, R.R.Salakhutdinov, Reducing the Dimensionality of Data with Neural Networks, Science, Vol.313, no. 5786, pp. 504–507, 2006.

５．主な発表論文等
〔雑誌論文〕（計 1 件）

① 李津, 實廣貴敏, 武田一哉, “単一マイクロホン入力音声から音響モデルを用いた発話者との距離推定,” 愛知工科大学紀要, 第 14 巻, pp. 1–7, 2017 年

〔学会発表〕(計 2 件)

① 宮嶋 博, 布目貴大, 實廣貴敏, 武田一哉, “Deep Belief Network を用いた単一マイクロホンによる発話者までの距離推定,” 日本音響学会 2018 年春季研究発表会講演論文集, 1-Q-1, pp. 61-62, 2018 年
② 井上拓哉, 實廣貴敏, “iBeacon を利用した音声対話システムの提案,” 日本音響学会 2015 年秋季研究発表会講演論文集, 3-Q-5, pp. 151-152, 2015 年

〔その他〕

① 宮嶋 博, “Deep Belief Network を用いた発話者距離認識における累積対数スコアによる精度向上の検討”, 愛知工科大学卒業論文, 2017 年
② 布目貴大, “Deep Belief Network による単一マイクロホン入力音声に対する距離認識の検討”, 愛知工科大学卒業論文, 2016 年
③ 加藤大貴, “Mel-Frequency Cepstrum Coefficients を用いた単一マイクロホンでの発話者距離推定”, 愛知工科大学卒業論文, 2015 年
④ 松井誉仁, “単一マイクロホンによる音響モデルを用いた発話者の実環境音声での評価”, 愛知工科大学卒業論文, 2014 年
⑤ ホームページ

愛知工科大学：實廣研究室：研究外部資金による研究
http://www1.aut.ac.jp/~jtlab/AUT_JTLAB/yan_jiu_zi_jin.html

６．研究組織

(1)研究代表者

實廣 貴敏 (JITSUHIRO, Takatoshi)
愛知工科大学・工学部・准教授
研究者番号: 60394996

(3)連携研究者

武田 一哉 (TAKEDA, Kazuya)
名古屋大学・大学院情報科学研究科・教授
研究者番号: 20273295

鹿野 清宏 (SHIKANO, Kiyohiro)
奈良先端科学技術大学院大学・名誉教授
研究者番号: 00263426