社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2008－SLP－74（5）
2008 / 12 / 9

音響情報と映像情報の統合による多人数会話における話者決定技術

石塚 健太郎† 荒木 章子† 大塚 和弘† 藤本 雅清† 中谷 智広†

†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

〒619-0237 京都府相楽郡精華町光台 2-4

E-mail: †{ishizuka, shoko}@cslab.kecl.ntt.co.jp, otsuka@eye.brl.ntt.co.jp, {masakiyo, nak}@cslab.kecl.ntt.co.jp

あらまし 本稿では，音響情報と映像情報を確率的に統合して用いることにより，多人数会話において「誰がいつ話したか」を推定する話者決定（Speaker Diarization）技術を提案する．音響情報と映像情報は，3 本のマイクロホンからなる三角形のマイクロホンアレイと魚眼レンズを装備した 2 台のカメラから構成される，多人数会話分析のための小規模システムを用いて収録される．このシステムで収録されたデータを元に話者決定を実現するために，提案技術は音声区間検出技術，話者方向推定技術，顔画像追跡技術から得られる発話の存在確率，話者の存在確率，会話参加者の存在確率を統合して用いる．日常的な雑談を用いた実験により，提案手法の話者決定における有効性が確認された．

キーワード 多人数会話分析，話者決定，マルチモーダル

Speaker diarization of multi-party conversations based on audio and visual information integration

Kentaro ISHIZUKA† Shoko ARAKI† Kazuhiro OTSUKA†

Masakiyo FUJIMOTO† and Tomohiro NAKATANI†

†NTT Communication Science Laboratories, NTT Corporation

2-4 Hikaridai, Seikacho, Sourakugun, Kyoto, 619-0237 Japan

E-mail: †{ishizuka, shoko}@cslab.kecl.ntt.co.jp, otsuka@eye.brl.ntt.co.jp, {masakiyo, nak}@cslab.kecl.ntt.co.jp

Abstract This paper proposes a speaker diarization method, which detects "who spoke when" in multi-party conversations, based on the probabilistic integration of audio and visual information. The audio and visual information is obtained from a compact system, which consists of two cameras with fisheye lenses and a triangular microphone array with three microphones, designed to analyze multi-party conversations. To realize speaker diarization, our proposed method utilizes the probability distributions of speech presence, speaker locations, and participants' face locations obtained with a speech activity detector, a direction of arrival based speaker location detector, and a face tracker, respectively. An experiment using real casual conversations revealed the advantages of such integration.

Keyword Multi-party conversation analysis, Speaker diarization, Multimodal systems

1. はじめに

会議など，多人数で行われる会話を収録し，自動的にインデクスを付与することができれば，蓄積された情報を事後に迅速に検索できたり，自動議事録作成を行ったり，自動要約することが可能になる[1][2][3][4]．このような多人数会話データの収録は，近年 Augmented Multi-party Interaction (AMI) [5], Computers in the Human Interaction Loop (CHIL) [6], NIST Rich Transcription Meeting Recognition Project [7]などのプロジェクトに関連して広く行われている．このような収録データに対する最も基本的なインデクスとして，会話中で「誰がいつ話したか」を推定する必要がある．この情報を推定するための技術は話者決定技術（Speaker diarization）と呼ばれ，多くの研究がなされている[8][9][10]．

近年，我々は多人数会話を分析するためのシステムを開発した[11]．このシステムは，魚眼レンズを持つ 2 つのカメラと，三角形に配置された 3 本のマイクロホ

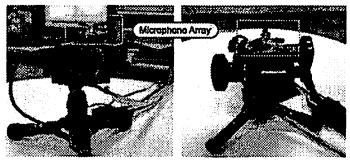

Figure 1: Omnidirectional camera-microphone system.

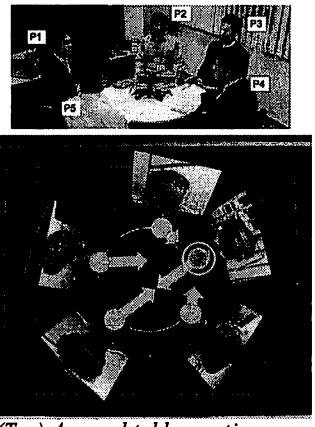

Figure 2: (Top) A round table meeting scene that can be analyzed with our system [11]. (Bottom) An example visualization result obtained with this system. The arrows show "who is looking at whom", and the red circle shows "who is speaking now".

ンから成るマイクロホンアレイの組み合わせにより構成されている．このシステムでは，雑音に頑健な音声区間検出（Speech Activity Detector; SAD）技術[12]，時間周波数領域で推定した信号の到来方向（Direction Of Arrival at each Time-Frequency bin; TFDOA）に基づく話者位置推定技術[13]，および Graphics Processing Unit (GPU)によるパーティクルフィルタリングを用いた顔の位置方向を推定する技術[14]を用いることで，会話中で「誰がいつ話したか」および「誰が誰を見ているか」を自動的に検出し可視化することができる．このシステムは円卓での会話を分析することを想定し，小規模なシステム構成になっており，会話中は円卓の中央に配置され，実時間処理により上記の情報を推定する．システム構成と会話状況の可視化の例を図 1 と図 2 にそれぞれ示す．

このシステムでは，上記 2 つの音響情報処理により「誰がいつ話しているか」を可視化することができるが，音響情報処理と映像情報処理がカスケード接続されていたことから，これまで顔位置追跡技術により推定される映像情報を音響情報処理のためには利用していなかった．本研究では，それらの要素間の情報を確率的に統合することを検討する．

近年，多人数会話を分析するために音響と映像の情報を統合する技術は広く研究されており，例えば，話者追跡[15][16]，グループ行動検出[17]，音声イベント検出[18]，アテンション認識[19]，参加者のインタラクション検出[20]などが，マルチモーダルな情報の統合により実現されている．これらに対し，本研究では，音響情報処理で得られる話者決定[14]の精度を，映像情報処理との併用により向上させることを目的とする．

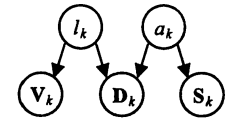

Figure 3: A graphical model that captures speaker locations and speech activities in group conversations. The arrows indicate the conditional dependencies between variables.

以後，本稿では，2 章で提案技術について述べ，3 章で実際の会話を用いた評価実験とその結果について述べ，4 章で結論を述べる．

2. 提案技術

話者決定は，観測される信号から「誰がいつ話したか」を自動検出する問題として定義される．円卓を対象とする上記のシステム構成に基づき，本稿では，会話中に話者が大きく移動しないことを仮定し，会話参加者の位置（方向角）が会話参加者そのものと対応するものとみなす．従って，本研究では，話者の位置と発話の有無を検出することで話者決定を実現する．

上記に基づき，本研究では方位角 k からの局所的な観測信号を用い，図 3 に示すグラフィカルモデルを考える．このモデルにおいて，l_k, a_k, $\mathbf{S}_k$, $\mathbf{D}_k$, $\mathbf{V}_k$ はそれぞれ話者の存在を表す命題，発話の存在を表す命題，方位角 k における音響信号の観測スペクトル，方位角 k における音響信号のパワー分布，方位角 k における観測映像信号を表す．これらに基づいて方位角 k ごとに事後確率 $p(a_k,l_k|\mathbf{S}_k,\mathbf{D}_k,\mathbf{V}_k)$を求め，この事後分布を閾値処理することで話者決定を実現する．この事後分布は $\mathbf{S}_k$, $\mathbf{D}_k$, $\mathbf{V}_k$ が相互に独立と仮定することで，以下のように記述できる．

$$p(a_k,l_k|\mathbf{S}_k,\mathbf{D}_k,\mathbf{V}_k)=p(a_k,l_k,\mathbf{S}_k,\mathbf{D}_k,\mathbf{V}_k)/p(\mathbf{S}_k)p(\mathbf{D}_k)p(\mathbf{V}_k) \quad (1)$$

式(1)の右辺はさらに以下のように記述できる．

$$p(a_k,l_k,\mathbf{S}_k,\mathbf{D}_k,\mathbf{V}_k)/p(\mathbf{S}_k)p(\mathbf{D}_k)p(\mathbf{V}_k)$$
$$=p(\mathbf{V}_k|a_k,l_k,\mathbf{S}_k,\mathbf{D}_k)p(a_k,l_k,\mathbf{S}_k,\mathbf{D}_k)/p(\mathbf{S}_k)p(\mathbf{D}_k)p(\mathbf{V}_k) \quad (2)$$
$$=p(\mathbf{V}_k|l_k)p(\mathbf{S}_k|a_k,l_k,\mathbf{D}_k)p(a_k,l_k|\mathbf{D}_k)/p(\mathbf{S}_k)p(\mathbf{V}_k) \quad (3)$$
$$=p(\mathbf{V}_k|l_k)p(\mathbf{S}_k|a_k)p(a_k,l_k|\mathbf{D}_k)/p(\mathbf{S}_k)p(\mathbf{V}_k) \quad (4)$$

ここで式(3)を求めるために $p(\mathbf{V}_k|a_k,l_k,\mathbf{S}_k,\mathbf{D}_k)=p(\mathbf{V}_k|l_k)$を仮定し，式(4)を求めるために $p(\mathbf{S}_k|a_k,l_k,\mathbf{D}_k)=p(\mathbf{S}_k|a_k)$を仮定した．さらに，式(4)の分母の第 1 項と第 2 項にベイズの定理を適用し，以下の式を得ることができる．

$$p(a_k, l_k | \mathbf{S}_k, \mathbf{D}_k, \mathbf{V}_k)$$
$$= p(l_k | \mathbf{V}_k) p(a_k | \mathbf{S}_k) p(a_k, l_k | \mathbf{D}_k) / p(a_k) p(l_k)$$
$$\propto p(l_k | \mathbf{V}_k) p(a_k | \mathbf{S}_k) p(a_k, l_k | \mathbf{D}_k) \quad (5)$$

ここで事前分布 $p(l_k)$と $p(a_k)$は定数と仮定した．

式(5)から，以下の 3 つの要素技術から推定される確率の乗算により，話者決定のための音響情報と映像情報の統合を行うことができることがわかる．まず，第 2 項の $p(a_k|\mathbf{S}_k)$は音響信号の観測スペクトルから求まる発話の存在確率であり，音声区間検出技術により推定できる．次に，第 3 項の $p(a_k,l_k|\mathbf{D}_k)$は音響信号の空間パワー分布から求まる話者の存在と発話の存在の同時確率であり，話者位置推定技術により推定できる．最後に，第 1 項の $p(l_k|\mathbf{V}_k)$は映像情報から得られる会話参加者の存在確率であり，顔追跡技術により推定することができる．このことから，会話分析システム[11]で既に利用している要素技術[12][13][14]から求まる確率を用いることで，音響情報と映像情報を統合した話者決定技術が実現可能であることがわかる．

以下，各要素技術において上記の各条件付確率を求める方法について述べる．なお，上記の事後確率 $p(a_k,l_k|\mathbf{S}_k,\mathbf{D}_k,\mathbf{V}_k)$は分析フレームごとに求める．

2.1. 音声区間検出技術

式(5)の条件付確率 $p(a_k|\mathbf{S}_k)$は，雑音に頑健な音声区間検出技術 MUlti Stream Combination of Likelihood Evolution for Voice Activity Detector (MUSCLE-VAD)により推定した．MUSCLE-VAD は，観測信号の周期性成分と非周期性成分の比（Periodic to Aperiodic component Ratio; PAR）に基づく雑音に頑健な音声区間検出法（PAR based Activity DEtection; PARADE）[22]と，Switching Kalman Filter（SKF）に基づく音声／非音声識別機構[23]を統合したものである（詳細は[12]参照）．本稿では，$p(a_k|\mathbf{S}_k)$の代わりに，方向性を持たない観測信号のスペクトル $\mathbf{S}$ から推定した $p(a_k|\mathbf{S})$を全ての方位角 k に対し用いることとする．上記 2 つの音声区間検出法ではそれぞれ，音声区間の検出を行う過程で観測 $\mathbf{S}$ に音声信号が含まれる仮説（H_1）と雑音信号のみが含まれる仮説（H_0）を考慮し，これらから得られる以下の尤度比を求め，尤度比検定により音声区間を推定している．

$$\Lambda = \frac{p(\mathbf{S}|H_1)}{p(\mathbf{S}|H_0)} \quad (6)$$

本研究では，この尤度比を用いることで，$p(a_k|\mathbf{S})$を推定する．

PAR は観測信号の周期性成分と非周期性成分のパワー比である．周期性成分の存在する周波数領域における非周期性成分のパワーを考慮した周期性・非周期性成分分解を行うことにより，その結果得られる PAR は，定常雑音だけでなく，突発性雑音を含む非定常の有色雑音に対しても頑健な性質を持つ．また，上記のパワー分解で推定される周期性・非周期性成分のパワーの推定誤差を考慮することで，観測信号中の音声の有無と対応する尤度比Λ(PAR)が推定される．

SKF に基づく音声／非音声決定機構で用いる音響特徴は 24 次のメルフィルタバンク出力である．この機構では，事前にクリーンな音声から音声／無音のガウス混合分布モデルを学習しておく．その一方で，雑音の性質については一般に事前に知ることが出来ないことから，雑音の周波数特性を，逐次状態が遷移するモデルによりモデル化し，非定常雑音に対処するため，カルマンフィルタによりその雑音のモデルを逐次更新する．上記 2 つの音声モデルと雑音モデルを合成して得られるモデルから，音声の有無と対応する尤度比Λ(SKF)を求めることができる．

最終的に，MUSCLE-VAD は上記 2 つの尤度比を乗算し，以下の尤度比Λを計算する．

$$\Lambda = \Lambda(PAR) \times \Lambda(SKF) \quad (7)$$

この尤度比Λを用いて以下の正規化を行い，条件付確率 $p(a_k|\mathbf{S}_k)$を求める．

$$p(a_k|\mathbf{S}) = \frac{\Lambda}{(1+\Lambda)} = \frac{p(\mathbf{S}|H_1)}{p(\mathbf{S}|H_1) + p(\mathbf{S}|H_0)} \quad (8)$$

2.2. 話者位置検出技術

式(5)の条件付確率 $p(a_k,l_k|\mathbf{D}_k)$は，観測信号の TFDOA に基づく話者位置検出技術によって推定した．この方法では，発話者 n の位置（方位角）に対応する重心 θ_1 .. θ_N を持つクラスタ C_1 .. C_N を推定する．クラスタ数と重心の位置は観測信号からオンラインで推定し，事前に話者数などを与える必要は無い（詳細は[13]参照）．これらの推定には，推定 DOA に基づいた Leader-follower クラスタリングと次節で述べる会話参加者の顔位置方向の検出結果を共に用いた．複数話者が同じ分析フレーム内で同時に発話している場合に対処するため，音声のスパース性を仮定し，TFDOA に基づく DOA 推定を行った．まず，信号の到来方向時間差（Time Difference Of Arrival; TDOA）を以下で求めた．

$$q'_{jj'}(f,\tau) = \frac{\arg\left(x_j(f,\tau)/x^*_{j'}(f,\tau)\right)}{2 \cdot \pi \cdot f} \quad (9)$$

ここで $x_j(f, \tau)$は j 番目のマイクロホンで観測された観測信号の分析フレーム τ, 周波数 f における離散フーリエ表現を表す．式(8)で求まる，各マイクロホンペアから求まる TDOA から TDOA ベクトル $\mathbf{q}'(f, \tau)$を構成し，これとマイクロホンの空間位置ベクトル $\mathbf{D}$ を用いることで，以下のようにして TFDOA ベクトル $\mathbf{q}(f, \tau)$を求めることができる．

$$\mathbf{q}(f,\tau) = c\mathbf{D}^+\mathbf{q}'(f,\tau) \quad (10)$$

ここで c は音速を表し，$^+$は Moore-Penrose 型の擬似逆行列を表す．TFDOA ベクトルから求まる各時間周波数における信号の到来方位角 $\theta^A(f, \tau)$から，条件付確率 $p(a_k, l_k|\mathbf{D}_k)$を以下のようにして求めることができる．

$$p(a_k, l_k|\mathbf{D}_k) = \begin{cases} \sum_f \phi_n(\theta^A(f,\tau))/F & \text{if } k \in C_n \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

ここで F は 1 分析フレームでの周波数ビンの総数を表し，ϕ は以下の操作を表す．

$$\phi_n(\theta^A(f,\tau)) = \begin{cases} 1 & \text{if } |\theta^A(f,\tau) - \theta_n^A| < threshold \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

ここで θ_n^Aは推定 DOA のクラスタリング[13]により求められた n 番目のクラスタの重心を表す．

2.3. 顔位置追跡技術

式(5)の条件付確率 $p(l_k|\mathbf{V}_k)$は，顔位置検出・追跡技術 Sparse Template Condensation Tracker（STCTracker）[14] により得られる顔位置方向情報から求めた．STCTracker は 3 次元の顔テンプレートを観測信号から自動的に生成し，顔が水平方向に 60 度回転した場合でも頑健に顔位置方向を追跡することが可能である．また，GPU を用いたパーティクルフィルタにより，実時間で複数人の顔位置方向を同時に追跡することができる．

図 1 に示した 2 つの魚眼レンズを装備した全方位カメラを用いることでほぼ全方位をカバーした映像を観測し，パノラマ変換により魚眼座標上における歪みを補正する．STCTracker の初期化処理ではスパースな特徴点を用いた疎テンプレート照合を行う．顔の位置と向きを表すテンプレートの状態は，画像平面を表す 2 自由度のベクトル，回転を表す 3 自由度のベクトル，スケール，および輝度から成る 7 次元のベクトルにより表現される．パーティクルフィルタ処理においては，パーティクル集合として表現されるテンプレートの状態の事後確率密度を逐次推定する．各パーティクルの重みは入力画像とテンプレートの照合誤差により推定される．上記のようにスパースな特徴点を用いること，およびパーティクルフィルタ処理による複数仮説の生成と検定による頑健なテンプレート照合により，STCTracker は実時間で頑健に動作する．顔モデルは対象人物毎に固定された形状をもつが，発話の生成や表情表出などによる顔の変形にも対処できる．図 2 では，複数の参加者の顔位置方向を実時間で同時に追跡できていることが，顔画像の上に表示されたメッシュにより示されている．

本研究では，STCTracker で求まる各話者の顔の中心の方位角 $\theta^V(\tau)$を会話参加者の位置の重心とみなして用いる．STCTraker の測定誤差から会話参加者の存在を表す条件付確率 $p(l_k|\mathbf{V}_k)$が求まると仮定し，この誤差

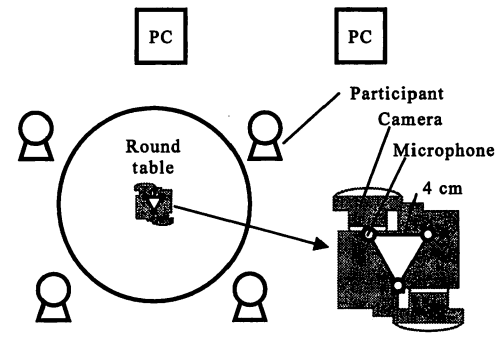

***Figure 4**: The experimental setup. Participants were all seated, and moved little during conversations.*

を以下に示す平均 $\theta_n^V(\tau)$，標準偏差 σのガウス分布によってモデル化する．

$$p(l_k|\mathbf{V}_k) = N(k; \theta_n^V(\tau), \sigma^2) \quad (13)$$

ここで $\theta_n^V(\tau)$は会話参加者 n の顔の方位角を示す．標準偏差 σは事前に固定の値（例えば 10 度）を与える．n の選択は k と $\theta_n^V(\tau)$が最小のものを選択することで行い，映像から顔が検出されていない場合は $p(l_k|\mathbf{V}_k)$を常に 0 とした．

3. 評価実験

3.1. 収録環境

データの収録は残響時間が約 350 ms の会議室で行った．図 4 に収録環境を示す．会話参加者 4 名（男女各 2 名）による 10 分の日常的な会話を 2 セット収録した．収録システムは円卓の中央に配置し，会話参加者は円卓を囲む形で座った状態で会話を行った．各会話参加者とマイクロホンアレイとの距離は約 1 m であった．収録されたデータの発話数や発話時間，2 人以上の同時発話があった同時発話時間の総数を表 1 に示す．

本収録システムは音響情報と映像情報の処理にそれぞれ 1 台の計算機（PC）を用いている．2 台の PC はデータ収録中に共に上記の会議室内に配置されていた．機材の構成の詳細は[11]と同様である．

音響信号のサンプリングレートは，音声区間検出技術では 16 kHz を用い，TFDOA の推定には 8 kHz にダウンサンプリングしたものを用いた．観測信号の分析フレーム長は 64 ms で，フレームシフトは 32 ms であった．2.2 節の話者クラスタリングに用いる閾値は 15 度とした．STCTracker はおよそ 24 フレーム/秒で処理を行い，各フレームで検出された顔位置方向を映像情報として用いた．2 台の PC 間では時間を同期させ，音響情報と映像情報の同期を取った．

なお，音響情報と映像情報からそれぞれ得られる方位角の情報を補正するため，以下の補正行列 $\mathbf{M}$ を推定した．$\mathbf{q}_n^V = [\cos\theta_n^V, \sin\theta_n^V]^T$ と $\mathbf{q}_n^A = [\cos\theta_n^A, \sin\theta_n^A]^T$ をそれぞれ n 番目の映像信号と音響信号から得られた方

Table 1: Data properties for evaluation experiment.

	Speaker ID	# utterances	Speech time (sec)	Overlap speech time (sec)
Conversation 1	1	170	200.7	171.1
	2	104	170.4	
	3	88	140.2	
	4	133	195.7	
Conversation 2	1	107	155.9	157.0
	2	113	195.6	
	3	60	102.0	
	4	175	237.2	

向ベクトルとし，θ_n^{V} と θ_n^{A} を音響情報と映像情報から得られた方位角とする．最小二乗誤差推定により，補正行列 **M** は，以下で求められる．

$$\sum_{n=1}^{N}\left\|\mathbf{q}_n^{\mathrm{A}}-\mathbf{M}\mathbf{q}_n^{\mathrm{V}}\right\|\xrightarrow[\mathbf{M}]{}\min. \qquad (14)$$
$$\mathbf{M}^{\mathrm{T}}=\left(\sum_{n=1}^{N}\mathbf{q}_n^{\mathrm{V}}\mathbf{q}_n^{\mathrm{V}\mathrm{T}}\right)^{-1}\left(\sum_{n=1}^{N}\mathbf{q}_n^{\mathrm{A}}\mathbf{q}_n^{\mathrm{V}\mathrm{T}}\right)^{\mathrm{T}}$$

正解ラベルは，収録されたデータに基づいてラベラーが各発話者の発話開始時刻と終了時刻を書き起こしたものを用いた．

3.2. 結果と考察

実験の評価尺度として，NIST Rich Transcription Meeting Recognition [7]で用いられている Diarization Error Rate (DER)を用いた．DER は，発話者のいない時間で発話を誤検出した時間 False-alarm Speech Time (FST), 発話者が存在する時間で発話無しとして誤棄却した時間 Missed Speech Time (MST), および話者の同定が誤っていた時間 Speaker Error Time (SET)の 3 種の誤り時間を合計し，それを総発話時間で除算することにより，以下のようにして求める．

$$DER=\frac{FST+MST+SET}{Total\ length\ of\ speech}\times 100\ \ (\%) \qquad (14)$$

発話区間の評価基準に関しても，NIST Rich Transcription [7]での基準を採用した．すなわち，発話区間は 300 ms 以上の無音区間で区切られるものとし，笑い声や咳などの voicing noise は無音区間の扱いとした．また，正解データに対する発話開始・終了時間のずれは 250 ms まで許容されることとした．

提案手法による音響情報と映像情報を統合した話者決定手法の有効性を評価するため，音響情報のみを用いた話者決定手法[13]との比較を行った．この方法では話者クラスタを生成するために過去一定時間内の音声フレーム数を用いている．

表 2 に表 1 に示した 2 つの収録データから得られた評価尺度を平均した実験結果を示す．今回収録されたデータは日常的な会話のデータであり，会議などよりも多くの話者交替，発話の重複，笑い声などが含まれていたため，話者決定が困難であった．実験の結果，音響情報と映像情報を統合することで，DER が 3.5 % 低下し，性能向上に繋がることがわかった．DER および MST の低下は，主に発話開始点を正確に検出できたことによるものであった．これは，STCTracker が音響事象の生じない状況（すなわち，発話の無い状況）であっても常に会話参加者の顔位置方向を追跡していたことによる効果と考えられる．このような正確な発話開始点検出は，多人数会話分析にとっても重要な役割を果たすと考えられる．その一方，提案手法においては，1 人の話者が発話している場合に，発話の無い他の会話参加者の発話を誤検出しやすくなる傾向があり，FST が従来手法よりも悪化した．これは音声の有無に対する感度が向上したことが原因と考えられる．

Table 2: Experimental results obtained with our proposed and previous methods [13].

	Proposed method			
	DER	MST	FST	SET
Proposed method	40.4	23.2	13.8	3.4
Previous method [13]	43.9	32.1	8.6	3.3 (%)

4. おわりに

本稿では，多人数会話を分析するための小規模システムにおける話者決定性能を向上させるための，確率的な音響情報と映像情報の統合について検討した．日常的な会話を用いた評価実験の結果，映像情報を用いることにより，音響情報単体で行うよりも話者決定の性能が向上することが分かった．

今後は，今回の検討を踏まえ，「誰がいつ話したか」を正確に推定するためのより効果的な音響・映像情報の統合を進める．それらに加え，あいづちや談話行為，話者の状態変化など，より上位の情報である会話参加者の振る舞いを分析する技術の研究や，自動話者認識／音声認識技術との併用を検討する予定である．

文　献

[1] A. Weibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, et al., "Advances in automatic meeting record creation and access," in Proc. Int. Conf. Acoust., Speech, and Signal Processing, 2001, pp. 597-600.

[2] N. Morgan, D. Baron, S. Bhagat, H. Carvey, R. Dhillon, J. Edwards, et al., "Meetings about meetings: Research at ICSI on speech multiparty conversations," in Int. Conf. Acoust., Speech, and Signal Processing, 2003, vol. 4, pp. 740-743.

[3] R. Cutler and L. Davis, "Distributed meetings: A meeting capture and broadcasting system," in Proc. ACM Int. Conf. Multimedia, 2002, pp. 503-512.

[4] Z. Yu, M. Ozeki, Y. Fujii, and Y. Nakamura, "Towards smart meeting: Enabling technologies and a real-world application," in Proc. ACM Int. Conf. Multimodal Interfaces, 2007, pp. 86-93.

[5] AMI and AMIDA Project: http://www.amiproject.org/

[6] CHIL Project: http://chil.server.de/

[7] NIST Rich Transcription Evaluation Project: http://nist.gov/speech/tests/rt/

[8] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," IEEE Trans. on Audio, Speech, and Language Processing, vol. 14, pp. 1557-1565, 2006.

[9] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," IEEE Trans. Audio, Speech, and Language Processing, vol. 15, pp. 2011-2022, 2007.

[10] D. Macho, J. Padrell, A. Abad, C. Nadeu, J. Hernando, J. McDonough, et al., "Automatic speech activity detection, source localization, and speech recognition on the CHIL seminar corpus," in Proc. Int. Conf. Multimedia Expo, 2005, pp. 876-879.

[11] K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, and J. Yamato, "A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization," in Proc. ACM Int. Conf. Multimodal Interfaces, 2008.

[12] M. Fujimoto, K. Ishizuka, and T. Nakatani, "A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme," in Proc. Int. Conf. Acoust., Speech, and Signal Processing, 2008, pp. 4441-4444.

[13] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "A DOA based speaker diarization system for real meetings," in Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays, 2008, pp. 29-32.

[14] O. Mateo Lozano and K. Otsuka, "Simultaneous and fast 3D tracking of multiple faces in video sequences by using a particle filter," J. Signal Processing Systems, DOI 10.1007/s11265-008-0250-2, in press.

[15] V. Khalidov, F. Forbes, M. Hansard, E. Arnaud, and R. Horaud, "Audio-visual clustering for 3D speaker localization," in Machine Learning for Multimodal Interaction, A. Popescu-Belis and R. Stiefelhagen (Eds.), LNCS vol. 5237, pp.86-97, 2008.

[16] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Audiovisual probabilistic tracking of multiple speakers in meetings," IEEE Trans. Audio, Speech, and Language Processing, vol. 15, pp. 601-616, 2007.

[17] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, pp. 305-317, 2005.

[18] F. Asano, K. Yamamoto, I. Hara, J. Ogata, T. Yoshimura, Y. Motomura, et al., "Detection and separation of speech event using audio and video information fusion and its application to robust speech interface," EURASIP J. Applied Signal Processing, vol. 11, pp. 1727-1738, 2004.

[19] S. O. Ba and J. M. Odobez, "Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues," in Proc. Int. Conf. Acoust., Speech, and Signal Processing, 2008, pp. 2221-2224.

[20] C. Busso, P. G. Georgiou, and S. S. Narayanan, "Real-time monitoring of participants' interaction in a meeting using audio-visual sensors," in Proc. Int. Conf. Acoust., Speech, and Signal Processing, 2007, vol. 2, pp. 685-688.

[21] G. Potamianos, J. Huang, E. Marcheret, V. Libal, R. Balchandran, M. Epstein, et al., "Far-field multimodal speech processing and conversational interaction in smart spaces," in Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays, 2008, pp. 119-123.

[22] K. Ishizuka and T. Nakatani, "Study of noise robust voice activity detection based on periodic component to aperiodic component ratio," in Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition, 2006, pp. 65-70.

[23] M. Fujimoto, K. Ishizuka, and T. Nakatani, "Noise robust voice activity detection based on switching Kalman filter," in Proc. Interspeech, 2007, pp. 2933-2936.