音声区間検出と方向情報を用いた会議音声話者識別システムとその評価*

○荒木章子, 藤本雅清, 石塚健太郎, 澤田宏, 牧野昭二 (NTT 研究所)

1 はじめに

本稿では、会議状況において「いつ誰が話したか」を、リアルタイムで推定する方法およびそのシステムについて説明する。提案法では、音声区間検出器 (VAD) で検出した音声区間における音声到来方向 (DOA) を分類することで、会議音声の話者識別を行う。本稿では、残響時間約 350ms の会議室にて収録した会議や会話音声の実録データについて、評価尺度 DER(Diarization Error Rate) により性能を評価した。その結果、話者交代や発話のオーバーラップが比較的多い会話音声についても、非常に小さい話者誤りで話者識別を行えることが示された。

2 背景

近年、NIST の Rich Transcription Meeting Recognition に代表されるように、多人数の会議を収録し、「いつ誰が話したか」を自動推定したり (diarization)、話者間のインタラクションを自動分析する研究が広く行なわれている (e.g., [1–5])。本稿では特に「いつ誰が話したか」を推定することを「話者識別」と呼ぶこととする。会議状況における話者識別は、会議データアノテーションやそれを用いた検索、会議録の自動作成、会議中の発言の音声強調 [6, 7] など、幅広い技術へ応用可能である。

ここで会議状況の定式化を行う。N 人の音声信号 $\mathrm{s}_1, \ldots, \mathrm{s}_N$ が、部屋の残響や雑音の影響を受け、M 個のマイクで観測されたとすると、観測信号は次のようにモデル化できる。

$$\mathrm{x}_j(t) = \sum_{k=1}^{N} \sum_l \mathrm{h}_{jk}(l)\, \mathrm{s}_k(t-l) + \mathrm{n}_j(t),\ j = 1, \ldots, M \quad (1)$$

ここで、$\mathrm{h}_{jk}(l)$ は音源 k からマイク j へのインパルス応答、$\mathrm{n}_j(t)$ はマイク j における雑音である。本稿では、音声信号 s_k は休止等を持つ自然な発話であり、収録中の席の移動は無いものとする。本稿の目的は、収録された観測信号 x_j のみから、フレーム毎に話者識別を行うことである。

本稿では、音声区間検出器 (VAD) で検出した音声区間について、音声到来方向 (DOA) を推定し、それを分類することで会議音声の話者識別を行う方法およびそのシステムを提案する。すなわち提案法は、話者の位置情報を手がかりとして話者識別を行う。本手法では、複数の VAD 手法を組み合わせ、多様な会議環境で想定される多様な雑音にも頑健性を持たせている。DOA を用いた話者分類は、これまでも [1] などで採用されているが、これらは多数の分散マイクを用いることを前提としている。一方、提案するシステムでは、3 個のマイクを正三角形に配置した小さなマイクアレイを用いることで、DOA と話者とを容易に対応づけることを可能とすると同時に、システムの可搬性を向上させている。

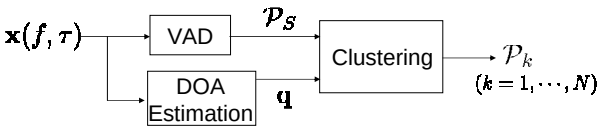

Fig. 1 Flow for proposed method.

本稿では、提案法およびシステムについて説明し、さらに、実際の会議室 (残響時間約 350ms) にて収録した会議/会話音声データを用いた性能評価の結果を報告する。

3 提案法

提案法のブロック図を図 1 に示す。本節では、図 1 のそれぞれのステップについて、詳しく説明する。

本稿では、時間周波数領域にて処理を行う。すなわち、式 (1) における観測信号 $\mathrm{x}_j(t)$ に短時間フーリエ変換 (short-time Fourier transform: STFT) を施し、時間周波数領域の観測信号 $x_j(f,\tau)$ を用いて処理を行う。ここで、f と τ はそれぞれ、周波数と、フレーム番号を示す。

3.1 音声区間検出 (VAD)

まず、音声区間検出器 (voice activity detector: VAD) を用いて、観測した会議音声の中から、音声区間を検出する。これは、雑音を誤ってある話者として識別することを防ぐためであり、発話の少ない会議や、方向性雑音のある環境において重要な役割を持つ。

本システムで用いた VAD のブロック図を、図 2 に示す。図に示すように、今回用いた VAD は、2 つの要素技術から構成されている。1 つは、信号の周期性成分と非周期性成分との比を用いた手法 (periodic to aperiodic component ratio-based detection ：PARADE) であり [8] 、もう 1 つは、確率モデルとスイッチングカルマンフィルタ (SKF) に基づく手法 [9] である。双方の VAD による音声/非音声の尤度について、その重みつき和を用いることで、最終的な音声区間 $\mathcal{P}_S$ を得る [10, 11]。

PARADE は、突発性雑音に対してロバストであるが調波構造を持つ雑音に弱い性質を持っており、一方 SKF は定常雑音に頑健であるが突発性雑音に弱い性質を持っていた。そこで、これらの 2 つの手法を統合

*A VAD-and-DOA-based meeting diarization system and its evaluation. by ARAKI, Shoko, FUJIMOTO, Masakiyo, ISHIZUKA, Kentaro, SAWADA, Hiroshi, MAKINO, Shoji (NTT Communication Science Laboratories, NTT Corporation)

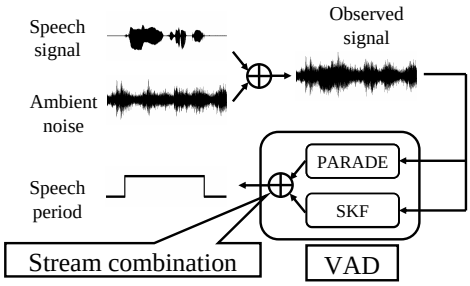

Fig. 2 Block diagram of VAD. PARADE: a Periodic to Aperiodic component RAtio-based DEtection, SKF: a switching Kalman filter.

することで、いろいろな種類の雑音に対して頑健な VAD を構築することが可能である [10, 11]。

PARADE [8] では、観測信号が、周期性成分 (基本周波数 F0 とその倍音成分から成る調波成分) と非周期性成分の和から成ると仮定する。各フレーム τ における周期性成分のパワー $\rho_p(\tau)$ と非周期性成分のパワー $\rho_a(\tau)$ との比 $\rho_p(\tau)/\rho_a(\tau)$(PAR) を特徴量として、各フレームの音声/非音声の尤度をそれぞれ求める。

SKF では、事前にクリーン音声データにて、クリーン音声と無音の GMM(Gaussian mixture model) を学習しておく。また雑音は状態遷移モデルで記述されると仮定し、カルマンフィルタにより、観測信号から雑音状態を逐次更新する。これらクリーン音声モデルと雑音モデルとを合成することで、雑音環境に適応した音声状態モデル (クリーン音声+雑音) と、非音声状態モデル (無音+雑音) の状態遷移モデルを生成する。そして、音声状態と非音声状態との尤度をそれぞれ計算する。ここで雑音モデルは逐次的に更新されることから、SKF では、雑音の時間変化に対して頑健な VAD を実現できる。

そして、最後に PARADE と SKF の結果を統合する。ここでは PARADE の尤度 $\gamma_{u,P}(\tau)$ と、SKF の尤度 $\gamma_{u,S}(\tau)$ (音声 $u=1$、非音声 $u=0$) を独立に計算した後、それぞれの尤度の重み付け加算を行い、最終的な尤度

$$\gamma_u(\tau) = (1-\lambda(\tau))\gamma_{u,P}(\tau) + \lambda(\tau)\gamma_{u,S}(\tau)$$

を得て、その尤度比 $\gamma_1(\tau)/\gamma_0(\tau)$ により音声区間 $\mathcal{P}_S$ を判定した [11]。本稿では、重み $\lambda(\tau)$ は固定値 0.8 とした。

本稿では、VAD の結果はバイナリラベルによって出力した。すなわち、非音声フレームは 0、音声フレームは 1 としてラベル付けした。また本稿では 3 個のマイクからなるマイクアレイを用いたがそれに対する VAD として、まず各チャネルに対して VAD を行い、その出力の論理和をとることで最終的なバイナリラベルを決定した。これにより、音声区間 $\mathcal{P}_S$ は、ラベルが 1 であるフレームの集合として決定した。

3.2 DOA 推定

次に、音声区間 $\mathcal{P}_S$ を、各話者区間 $\mathcal{P}_k$ ($k = 1, \cdots, N$) に分類することで、話者識別を行う。ここで、$\mathcal{P}_S = \sum_{k=1}^{N} \mathcal{P}_k$ とした。

本稿では、話者の特徴量として、音声区間 $\tau \in \mathcal{P}_S$ における、音声の到来方向 (direction of arrival: DOA) ベクトル $\mathbf{q}(\tau)$ を用いた。そして、推定 DOA がある一定範囲の値を取るフレーム τ の区間を同一話者区間 $\mathcal{P}_k$ とした。

各フレームにおける DOA 推定法は以下である。はじめに GCC-PHAT 法 [12] を用いて、全てのマイクペア jj' に関して音声の到来時間差 (time differences of arrival: TDOA) $q'_{jj'}(\tau)$ を推定する：

$$q'_{jj'}(\tau) = \mathrm{argmax}_{q'} \sum_f \frac{x_j(f,\tau)x^*_{j'}(f,\tau)}{|x_j(f,\tau)x^*_{j'}(f,\tau)|} e^{\jmath 2\pi f q'} \tag{2}$$

そして、全てのマイクペアにおける TDOA 値を並べたベクトルを $\mathbf{q}'(\tau)$ とする。

DOA ベクトル $\mathbf{q}(\tau)$ は、TDOA の推定値 $\mathbf{q}'(\tau)$ およびマイク座標を表す行列 $\mathbf{D}$ より

$$\mathbf{q}(\tau) = c\mathbf{D}^{-}\mathbf{q}'(\tau), \quad \mathbf{q}(\tau) \leftarrow \mathbf{q}(\tau)/||\mathbf{q}(\tau)|| \tag{3}$$

にて推定できる [13]。ここで c は音速、$^-$ は一般化逆行列である。尚、DOA ベクトル $\mathbf{q}(\tau)$ は、音源の方位角を $\theta(\tau)$、仰角を $\phi(\tau)$ とすると、$\mathbf{q}(\tau) = [\cos\theta(\tau)\cos\phi(\tau), \sin\theta(\tau)\cos\phi(\tau), \sin\phi(\tau)]^T$ であり、これより音声が到来する方位角と仰角を推定できる。

ここでは TDOA 推定に GCC-PHAT を採用したため、特徴量である DOA ベクトル $\mathbf{q}(\tau)$ は各フレームにつき 1 つだけ推定される (各時間周波数にて推定されるものではない)。尚、TDOA ベクトル $\mathbf{q}'(\tau)$ を特徴量として用いることも可能ではあるが、今回は、結果表示の直感的理解のしやすさを鑑み、DOA ベクトル $\mathbf{q}(\tau)$ を特徴量として用いた。

3.3 クラスタリング

次に、音声区間 $\mathcal{P}_S$ を各話者の発話区間 $\mathcal{P}_k$ に分類するため、音声区間 $\tau \in \mathcal{P}_S$ における特徴量 $\mathbf{q}(\tau)$ をクラスタリングする。ここでは、話者数 N が未知の場合にもクラスタリング可能とするために、leader-follower クラスタリングによるオンラインクラスタリングアルゴリズムを用いた [14]。この方法では、新たな話者が収録データに現れた時に、新たなセントロイドを生成し、クラスタリングを行う。

クラスタリングで得られる各クラスタが、各話者に対応しており、各話者の発話区間 $\mathcal{P}_k$ は、

$$\tau \in \mathcal{P}_k \quad \text{if} \quad \mathbf{q}(\tau) \in C_k \tag{4}$$

として求める。ここで、C_k は k 番目のクラスタである。

4 評価実験とその結果

4.1 実験条件

図 3 に示す室内にて収録した会議および会話音声を用いて評価実験を行った。部屋の残響時間はおよそ 350ms であった。各会議や会話は 3 名か 4 名の出席者にて行われ、1 つの収録の間、席の移動は無いものとした。各話者とマイクアレイの距離はおよそ 1m であった。部屋には、パソコンが 2 台（PC1, 2）と

Table 1　Conversation recordings. Each recording duration was five minutes.

Evaluation data ID	#Speaker	Overlap [%]	#Turn-taking	#Utterance	Noise sources
PR1 (presentation rehearsal 1)	4	1.4	40	119	PC1,2, projector, laughing voice
PR2 (presentation rehearsal 2)	3	6.0	75	145	of other speakers
CO1 (conversation)	3	34.8	243	278	PC2
DI1 (discussion)	3	10.8	126	172	PC2, paper noise
CP1 (crossword puzzle 1)	4	18.6	149	185	PC2
CP2 (crossword puzzle 2)	4	13.0	183	218	PC2

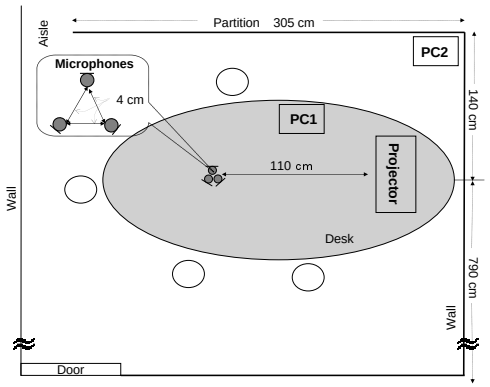

Fig. 3　Room setup. Small ellipses illustrate example speaker places.

プロジェクタが 1 台設置されており (図 3)、これらが稼動しているときは雑音源となった。

表 1 に、各収録の条件等を示す。今回収録した音声は会話的・放談的なものが多く、フォーマルな会議音声に比べ、録音音声に含まれる話者交代や話者音声のオーバーラップの頻度が多くなっている。そのために、今回の収録音声の話者識別タスクは、比較的難しいものとなっている。

話者識別の正解データとしては、各収録音声について、各話者の発話開始時刻と発話終了時刻を、手作業で付与したものを用いた。サンプリング周波数は 16kHz、STFT フレーム長は 64 ms、フレームシフトは 32 ms とした。

4.2　VAD の評価

ここでは、本システムで用いた VAD の性能を、雑音に頑健とされている Sohn らの手法 [15] と比較した。Sohn らの手法では、MMSE 推定スペクトルから求まる推定 SN 比を特徴量とし、観測信号が音声状態/非音声状態それぞれに属する尤度比を用いて、音声/非音声の識別を行う。

PARADE と SKF の特徴量としては、それぞれ、1 次元の周期成分対非周期成分比 (PAR) と、24 次の対数メルスペクトルを用いた。クリーン音声と無音の GMM の学習は、日本語音素バランス文 (101 話者、5050 文) を用いて行い、GMM の混合分布数は双方とも 32 とした。

評価尺度としては、false acceptance rate (FAR) と false rejection rate (FRR) を用いた:

$$\mathrm{FAR} = N_{FA}/N_{ns} \times 100\ [\%]$$
$$\mathrm{FRR} = N_{FR}/N_{s} \times 100\ [\%]$$

ここで N_{ns}, N_s, N_{FA}, N_{FR} はそれぞれ、非音声フレーム数、音声フレーム数、非音声を音声と誤検出したフレーム数、および音声を非音声と誤検出したフレーム数である。加えてここでは、NIST にて提案されている VAD の diarization error rate (DER) [3] も評価した:

$$\mathrm{DER} = \frac{\text{誤受理・誤識別した時間長}}{\text{全音声区間長}} \times 100[\%]$$

DER の測定基準についても、NIST 基準に準拠した。すなわち音声信号区間は 300ms 以上の非音声区間で区切られ、笑い声・咳などは非音声として扱い、発話区間の開始終了時刻の推定値は、正解ラベルに対し前後 250ms までのずれを許容した。

Table 2　Experimental results of VAD [%]

Data ID	Sohn				Proposed			
	FAR	FRR	Ave.	DER	FAR	FRR	Ave.	DER
PR1	41.8	14.3	28.1	20.0	12.8	24.3	**18.6**	22.2
PR2	24.3	28.5	26.4	34.2	22.5	19.7	**21.1**	**22.0**
CO1	33.2	44.1	38.7	56.6	47.9	21.7	**34.8**	**32.5**
DI1	56.4	17.0	36.7	45.7	14.8	22.4	**18.6**	**23.0**
CP1	12.6	37.1	24.9	45.3	16.4	13.7	**15.1**	**19.4**
CP2	32.0	22.8	27.4	37.3	26.8	13.6	**20.2**	**17.8**

表 2 に VAD の評価結果を示す。表より、今回採用した提案法は、FAR、FRR、DER のすべての尺度において、Sohn らの方法よりも高い性能を持つことがわかる。これは、環境に応じて適応的に雑音モデルを学習できる SKF の枠組みと、紙雑音などの突発的な雑音に強い PARADE とが相補的にうまく働き、会議環境における多様な雑音への頑健性が高まったためであると考えられる。特に、実際の会議状況では、定常・非定常とも様々な雑音の存在が考えられるため、このような相補的な働きのある VAD を組み合わせて用いることは効果的である。

4.3　話者識別の評価

ここでも評価指標としては、NIST によって提案された diarization error rate (DER) を用いた [3]。話者識別における DER は以下で定義される。

$$\mathrm{DER} = \frac{\text{誤受理・誤棄却・話者誤りの時間長}}{\text{全音声区間長}} \times 100[\%],$$

すなわち DER は、誤棄却 (missed speaker time: MST), 誤受理 (false alarm speaker time: FAT), 話者誤り (speaker error time: SET) の 3 つの誤検出を含む指標となっている。評価に際しては、(4) にて得られた $\mathcal{P}_k$ を時間方向にスムージングし (Hangover)、数フレーム以下の発話や無音区間を取り除いた後に、各話者の発話開始時刻と発話終了時刻を判定した。正解ラベルにおける話者と推定における話者は、発話方向を手がかりに対応づけた。もし、推定された話者数が、実際に会議に参加した話者数より多い場合、多く判定された話者は「ゴースト」と判定し、SET

Table 3　Experimental results of diarization [%]

Data ID	With Sohn's VAD				With proposed VAD			
	DER	MST	FAT	SET	DER	MST	FAT	SET
PR1	27.2	21.6	4.0	1.7	**23.9**	20.7	3.1	0.1
PR2	35.0	27.6	4.5	2.9	**31.2**	23.1	5.6	2.5
CO1	61.3	30.6	13.7	17.1	**38.7**	19.0	14.5	5.3
DI1	45.0	24.5	17.7	2.8	**34.8**	25.9	6.9	2.0
CP1	45.0	30.7	7.7	6.6	**36.9**	18.1	13.6	5.2
CP2	47.6	34.3	9.2	4.1	**32.7**	21.3	6.8	4.6

として評価した。その他の評価基準は 4.2 節と同様である。

表 3 に話者識別の結果を示す。提案の VAD を用いることで、Sohn らの VAD を用いた場合より低い DER に押さえることができることが分かった。また、今回提案の話者識別方法では、話者誤り (SET) が小さく押さえられることが分かった。これは、今回のような席を固定した会議/会話状況においては、方向情報が話者識別に有用であることを示している。また今回は、誤棄却 (MST) が多かった。この理由としては、各フレームで DOA を 1 つしか推定しない方法を採用したため、同一フレームにおける話者オーバーラップを拾いきれなかったためであると考える。これに対しては、各フレームで複数の DOA 値を出力する方法 (e.g.,[13]) を用いることで、MST の改善が見込めると考えている。

4.4　システム

構築したシステムは、マイクアレイで観測した音声を A/D 変換器にて PC に取り込み、本稿で述べた VAD と話者識別を行って、図 4 に例示する話者識別結果を表示するものである。ここで、図 4(a) にはマイク 1 における観測信号波形を、図 4(B) および (C) には、話者識別結果 $\mathcal{P}_k$ (式 (4)) をそれぞれ表示する。本システムは、PC1 台 (AMD Athlon64, 2.4GHz) でほぼリアルタイムで動作する。実装においては、VAD の部分は C 言語で、話者識別の部分および描画については Matlab6.5 で、それぞれ構築しており、システムのリアルタイムファクター (処理時間/データ長) はおよそ 0.6 であった。

5　まとめ

本稿では、会議状況において「いつ誰が話したか」を推定する方法について述べ、またリアルタイムにてそれを推定するシステムについて紹介した。提案法では、音声区間検出器 (VAD) で検出した音声区間についての音声到来方向 (DOA) を分類することにより、会議音声の話者識別を行った。実験より、DOA の分類により話者誤りの少ない話者識別ができること、雑音に頑健な VAD を採用することにより DER(Diarization Error Rate) が改善することが分かった。

尚、本稿では触れなかったが、「いつ誰が話したか」の情報を用いた音声強調も可能である。例えば、聞きたい話者音声のパワーと、その他の話者音声のパワーの比 (SN 比) を最大化するフィルタを設計することで、音声強調をすることができる [6, 7, 16]。これは例えば、雑音や発話オーバーラップの多い会議音声を後から聴取する場合などに有用である。

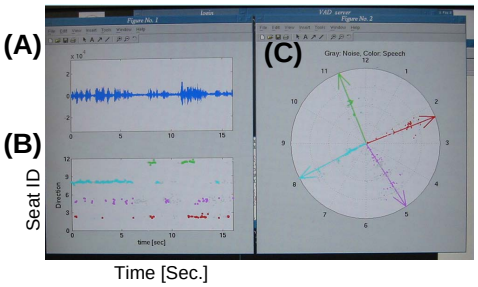

Fig. 4　Result image. (A) recording at microphone 1, (B) speaker indexing result, (C) speaker positions with respect to the microphone array (the center of the circle indicates the array position).

参考文献

[1] J. M. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multi-microphone meetings using only between-channel differences," in *Proc. of MLMI'06 (LNCS 4299)*. 2006, pp. 257–264, Springer.

[2] 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.

[3] `http://www.nist.gov/speech/test_beds/mr_proj/`.

[4] D. Ellis and J. Liu, "Speaker turn segmentation based on between-channel differences," in *Proc. of NIST Meeting Recognition Workshop*, 2004, pp. 112–117.

[5] C. Busso, P. Panayiotis, G. Georgiou, and S. Narayanan, "Real-time monitoring of participants' interaction in a meeting using audio-visual sensors," in *Proc. of ICASSP'07*, 2007, vol. II, pp. 685–688.

[6] S. Araki, H. Sawada, and S. Makino, "Blind speech separation in a meeting situation with maximum SNR beamformers," in *Proc. of ICASSP'07*, 2007, vol. I, pp. 41–45.

[7] 荒木章子, 澤田宏, 牧野昭二, "話者分類と SN 比最大化ビームフォーマに基づく会議音声強調," 音講論 (春), pp. 571–572, 2007.

[8] K. Ishizuka, T. Nakatani, M. Fujimoto, and N. Miyazaki, "Noise robust front-end processing with voice activity detection based on periodic to aperiodic component ratio," in *Proc. of Interspeech '07*, 2007, pp. 230–233.

[9] M. Fujimoto and K. Ishizuka, "Noise robust voice activity detection based on switching Kalman filter," in *Proc. of Interspeech '07*, 2007, pp. 2933–2936.

[10] M. Fujimoto, K. Ishizuka, and T. Nakatani, "A voice activity detection based on adaptive integration of multiple speech feature and signal decision scheme," in *Proc. of ICASSP '08*, 2008, (to appear).

[11] 藤本 雅清, 石塚健太郎, 中谷 智広, "複数の音声区間検出法の適応的統合の検討と考察," 電子情報通信学会, 音声研究会, SP2007-97, pp. 7-12, 2007.

[12] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust. Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[13] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *Proc. of ICASSP'06*, 2006, vol. 5, pp. 33–36.

[14] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley Interscience, 2nd edition, 2000.

[15] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[16] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada and S. Makino, "Speaker indexing and speech enhancement in real meetings / conversations," in *Proc. of ICASSP'08*, 2008, (to appear).