NTT

トピックモデルの応用：音声・音響データ

NTT コミュニケーション科学基礎研究所
石黒　勝彦

2013/01/15-16 統計数理研究所　会議室1

このスライドの“トピック”

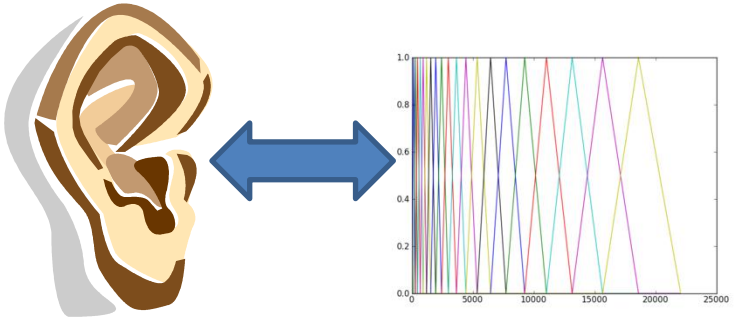
- 画像認識系から少し遅れますが、最近では音声・音響データに対してもトピックモデルが利用されるようになっています

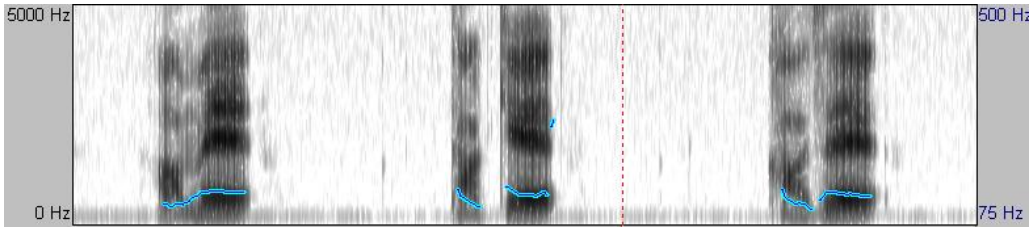
音声・音響信号で考えるべき問題

- 1. どの特徴量を利用するか？
- 2. 時系列性をどう扱うか？

音声・音響信号からは多彩な特徴量が抽出できます

- どの特徴量を利用して、どうやってBoW形式に変換するかを検討する必要があります
  - MFCC: 音声認識などで広い範囲で利用される
  - F0: 発話のイントネーションやメロディを表現





MFCC: 人間の
音声知覚を反映した(とされる)特徴

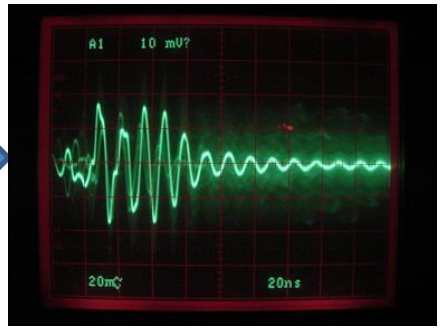
F0: 波形の基本周波数。ピッチ。

http://d.hatena.ne.jp/aidiary/20120225/1330179868
https://sites.google.com/site/utsakr/Home/courses/tsukuba2010/contents/13-f0

音声・音響信号は
複雑な時系列信号です

- マルコフ性を仮定する時系列モデルを利用するのが王道ですが、その必要があるかどうかの検討も必要です

$$f(t) = \int g(t-\tau)h(\tau)d\tau$$



http://plaza.rakuten.co.jp/DFLYRoom/diary/200808030000/

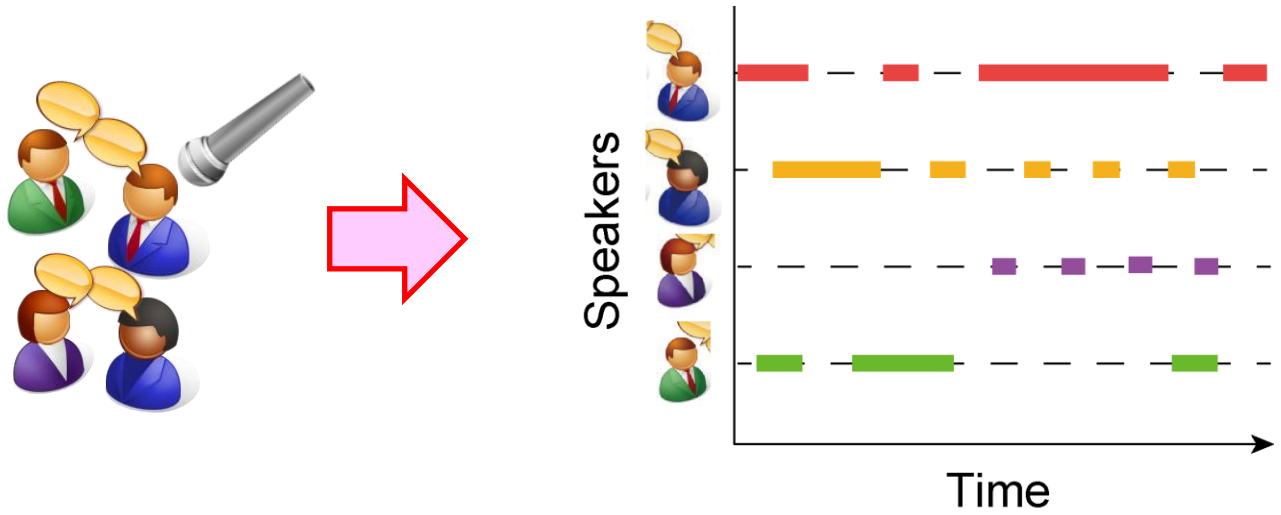
Topic Model for speaker diarization

[Ishiguro, 2012]

Ishiguro et al. ,
"Probabilistic Speaker Diarization with Bag-of-Words Representations of Speaker Angle Information",
IEEE Trans. ASLP, Vol. 20(2), pp. 447-460, 2012.

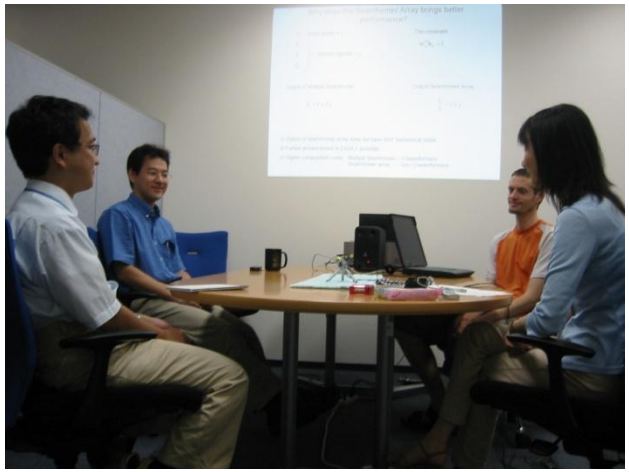
speaker diarization
“誰がいつ発話したか”

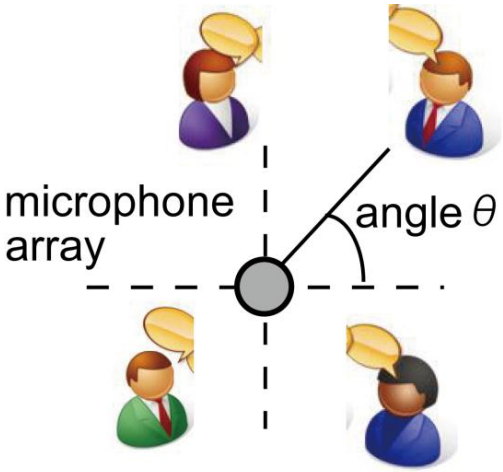
- 複数の音源があるときに、各音源がいつ信号を発信したかを決定
- 応用範囲: 会議の自動議事録作成、テレビ電話における発話者音声強調、ロボットと人間のインタラクションなど



会議状況のdiarization

- テーブルにマイクを置いて、会議状況をdiarizationします
- 一般に何人の話者がどこに座るかは事前にわかりません → 話者は潜在的な隠れ要素です
- その時々によって発話者が代わります → 各話者の発話状況は時間変化します



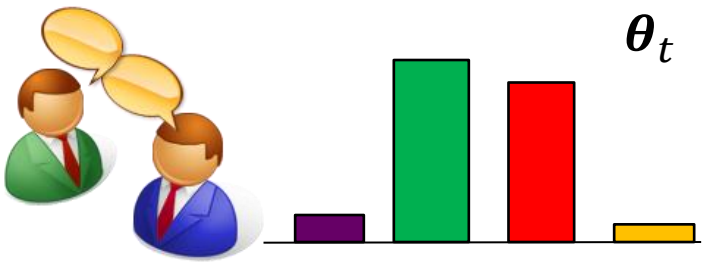


トピックモデルによるdiarization

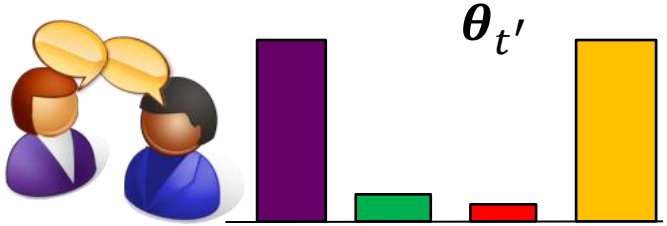
☺ diarizationタスクは自然にトピックモデルで記述できます

- 時刻 ＝ 文書と考えると、各時刻の発話は複数の潜在トピック ＝ 話者で表現できます
- トピック（話者）はわからないので推定します
- トピック分布に発話状況が反映されます

時刻 t



時刻 t′



提案法の概要

- diarizationとトピックモデルの共通点に気付いたことで、「話者＝トピック」と「各時刻の発話状態＝文書のトピック分布」を同時に推定できます

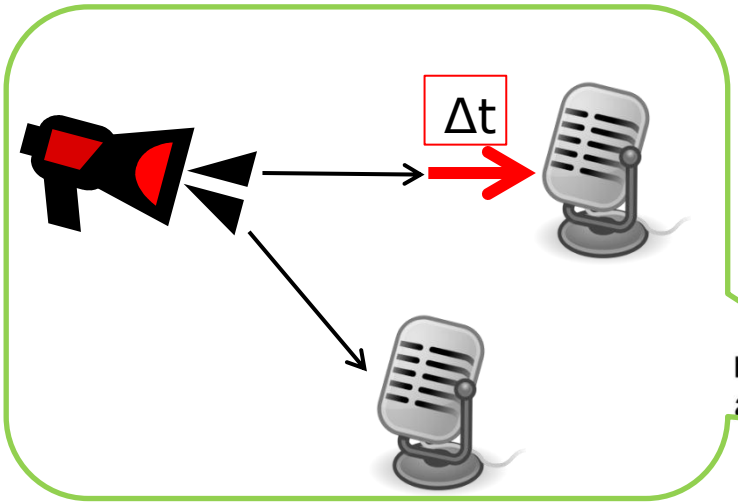
- diarizationに対するベイジアンモデルを提案できます

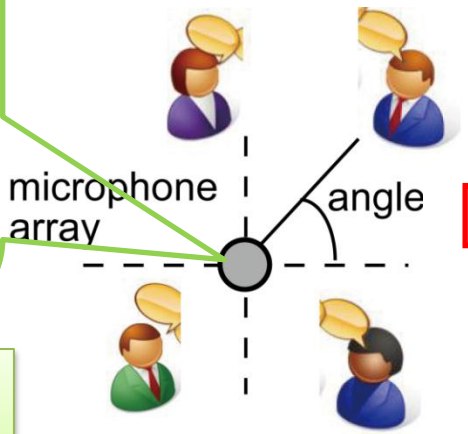
提案法のアイデア

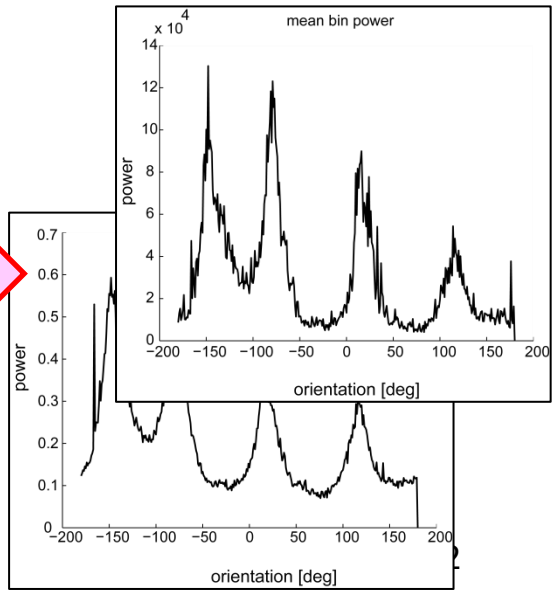
- 考えるべき2つの問題に以下のように対応します
- 特徴量: 方向情報 (DOA) → Bag of Angle Words
- 時系列性: 非定常な話者分布変化 → トピック分布の線形補間モデル

話者はどこにいるのか: DOAクラスタリング [cf. Araki, 2008]

- DOA: 音の聞こえてくる方向の特徴量
- クラスタリングによって、「話者がどこにいるのか」を推定できることが分かっています



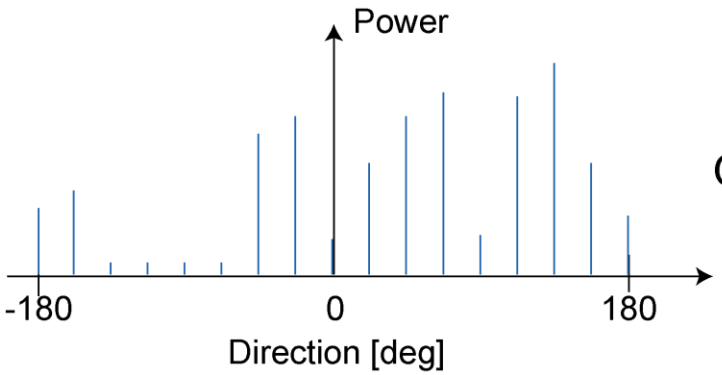




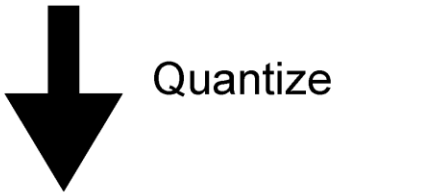
信号の受信時間差とマイク配置から音源位置が計算できる

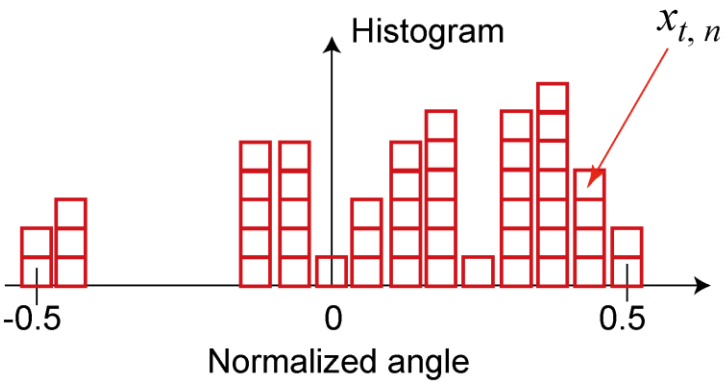
特徴量: Bag of Angle Words

☺ DOA特徴量を離散化、トピックモデルに使えるようにします



Observed DOA power





Bag of Angle words

[Ishiguro, 2012]

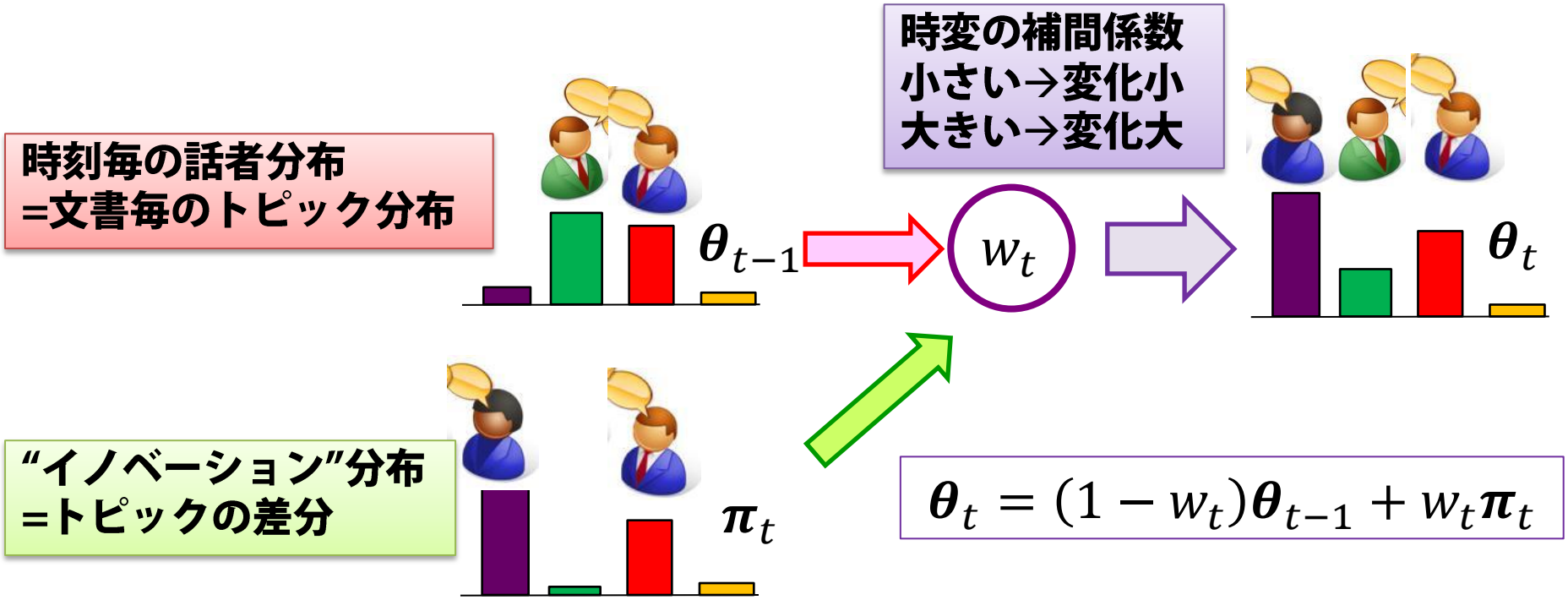
非定常な話者分布変化

- 時間連続性: ミリ秒単位の時間ステップでは、話者の発話分布は変わりません
- 時間非連続性: 発言を受けての応答など、会議の流れにそって話者分布が変化します (turn-taking)

- つまり、話者の発話状態の変化自体が非定常になっています

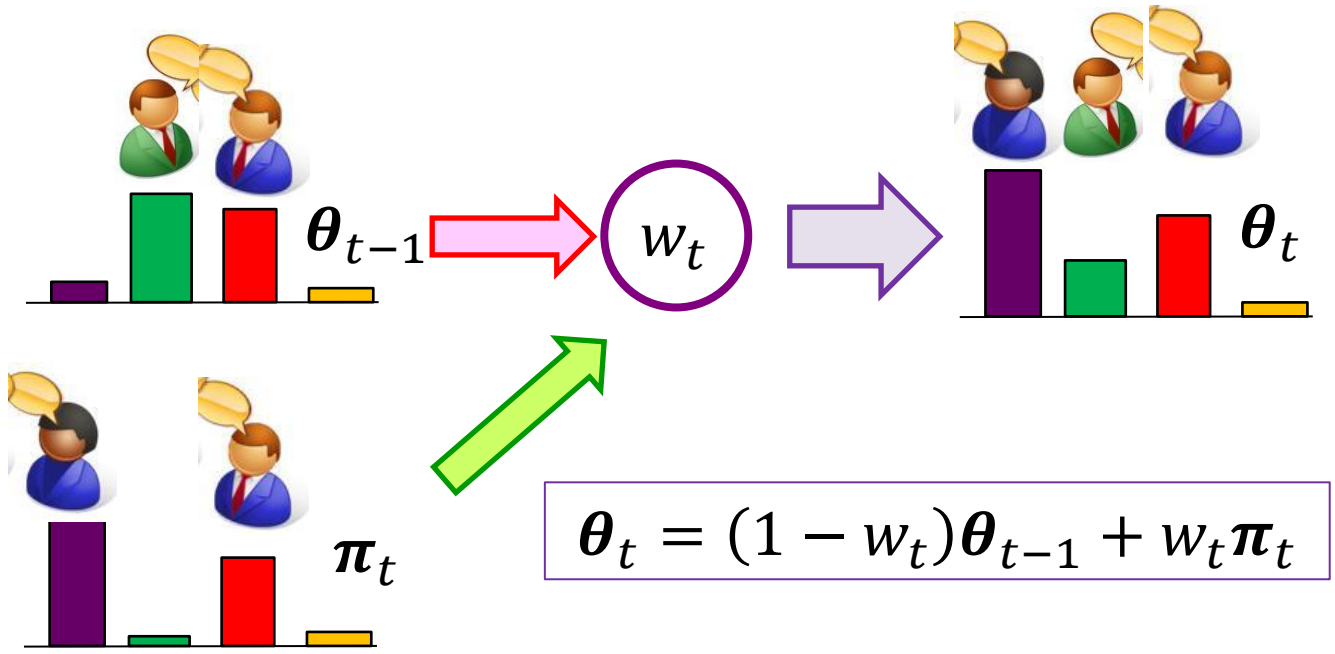
時系列性: トピック分布の線形補間モデル

- 話者分布の時間変化の非定常性を表すために、時変の補間係数を導入します



時系列性: トピック分布の線形補間モデル

- ☺簡単な線形モデルによるLDAの時間発展モデル
- ☺小規模～大幅な話者変化を表現可能
  - 前時刻との依存度をw_tで制御する



相互に独立なモデルへの変換

- 各時刻の話者分布θは、時刻ごとに独立なπの組み合わせで表現できます
- ☺ マルコフ性が消えて推論が簡単になります

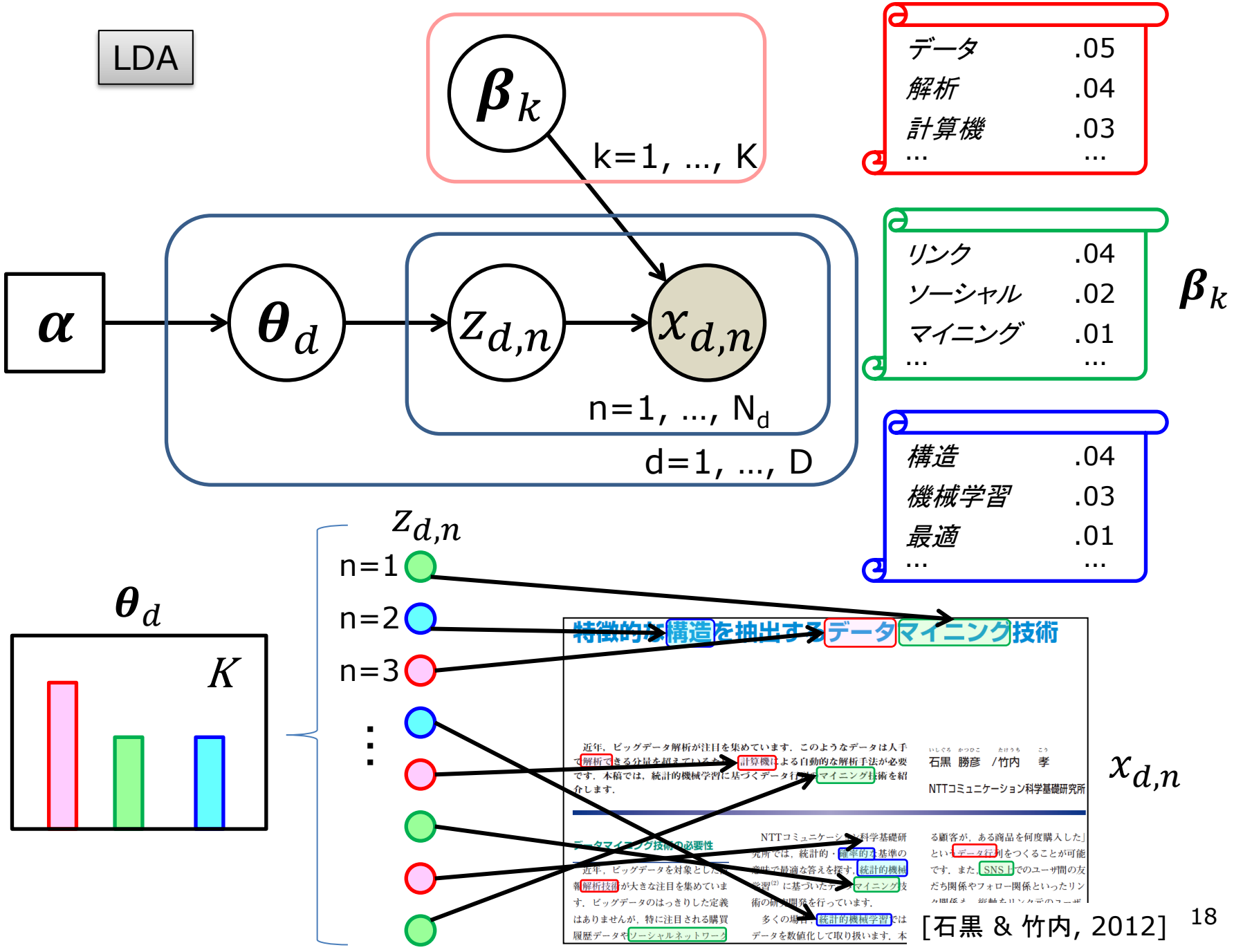
$$\boldsymbol{\theta}_t = (1 - w_t)\boldsymbol{\theta}_{t-1} + w_t\boldsymbol{\pi}_t$$

$$= (1 - w_t)\{(1 - w_{t-1})\boldsymbol{\theta}_{t-2} + w_{t-1}\boldsymbol{\pi}_{t-1}\} + w_t\boldsymbol{\pi}_t$$

$$\cdots$$

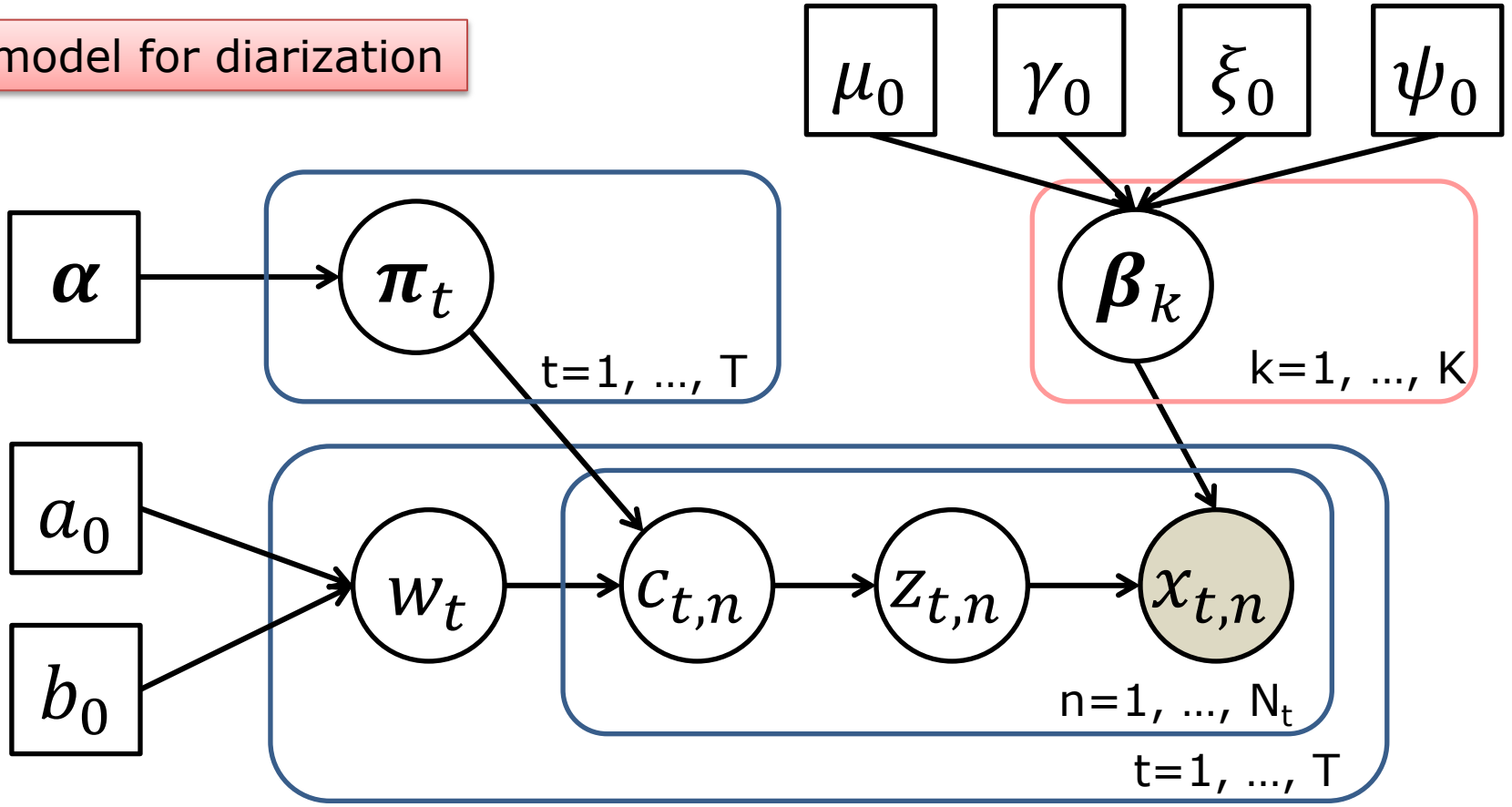
$$= \sum_{l=1}^{t} v_{tl}\boldsymbol{\pi}_l \qquad v_{tl} = w_l \prod_{m=l+1}^{t} (1 - w_m)$$

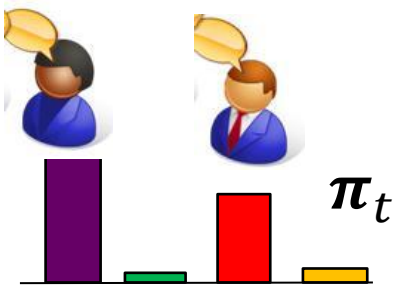
LDA



[石黒 & 竹内, 2012]

Topic model for diarization

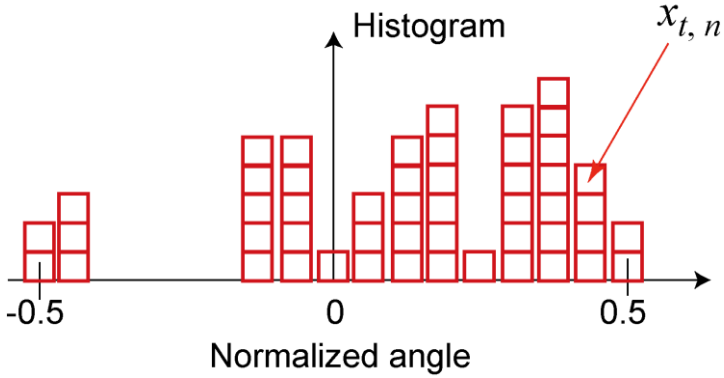




$$\boldsymbol{\theta}_t = (1 - w_t)\boldsymbol{\theta}_{t-1} + w_t\boldsymbol{\pi}_t$$

$$= \sum_{l=1}^{t} v_{tl}\boldsymbol{\pi}_l$$

$$v_{tl} = w_l \prod_{m=l+1}^{t} (1 - w_m)$$



生成モデル

for 時間t = 1, 2, …, T

"innovation" topic proportion $\boldsymbol{\pi}_t|\boldsymbol{\alpha}\sim\mathrm{Dir}(\boldsymbol{\alpha})$

interpolation factor $w_t|a_0,b_0\sim\mathrm{Beta}(a_0,b_0)$

for l = 1, 2, …, t

$$v_{tl}=w_l\prod_{m=l+1}^{t}(1-w_m)$$

for 単語n = 1, 2, …, $N_{t,d}$

for speaker (topic) k = 1, 2, …, K

topic-Angle word proportion

$$\boldsymbol{\beta}_k|\mu_0,\gamma_0,\xi_0,\psi_0\sim\mathrm{NormalGamma}(\mu_0,\gamma_0,\xi_0,\psi_0)$$

生成モデル

for 時間t = 1, 2, …, T

$$\boldsymbol{\pi}_t|\boldsymbol{\alpha}\sim\mathrm{Dirichlet}(\boldsymbol{\alpha}) \qquad v_{tl} = w_l \prod_{m=l+1}^{t} (1 - w_m)$$

for 単語n = 1, 2, …, $N_{t,d}$

innovation topic dist.-word assignment

$$c_{t,n}|\boldsymbol{v}_t\sim\mathrm{Mult}(\boldsymbol{v}_t)$$

speaker-angle word assignment

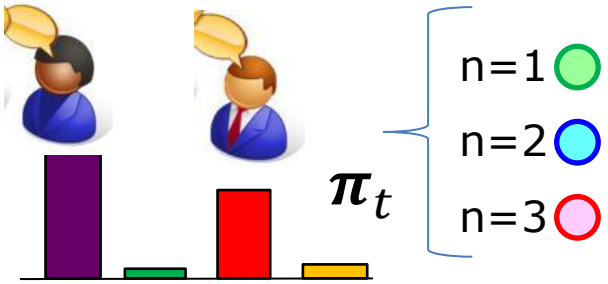
$$z_{t,n}|c_{t,n},\{\boldsymbol{\pi}_t\}\sim\mathrm{Mult}\left(\boldsymbol{\pi}_{c_{t,n}}\right)$$

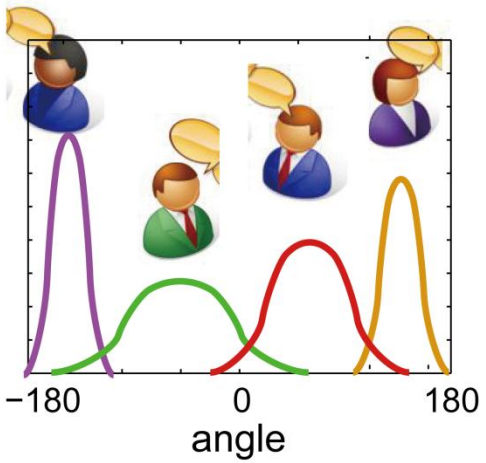
Angle word observation

$$x_{t,n}|z_{t,n},\left\{\boldsymbol{\beta}_{t,k}\right\}\sim\mathrm{N}\left(\boldsymbol{\beta}_{t,z_{d,n}}\right)$$

混合分布による疑似的なBag of Angle Words

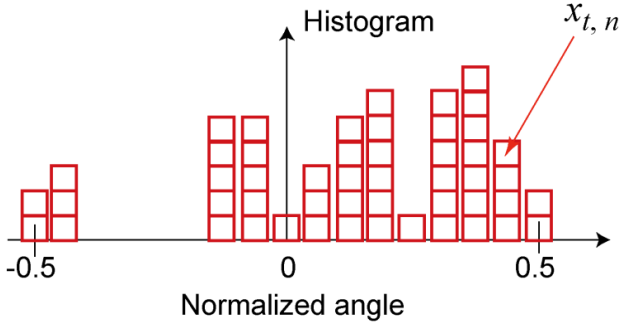
- Angle wordの値（角度・位置）には意味があるのでNormalから生成します





$$z_{t,n}|c_{t,n},\{\boldsymbol{\pi}_t\}\sim\mathrm{Multi}\left(\boldsymbol{\pi}_{c_{t,n}}\right) \qquad \boldsymbol{\beta}_k=\{\mu_k,\sigma^2\}$$

$$x_{t,n}|z_{t,n},\{\boldsymbol{\beta}_{t,k}\}\sim\mathrm{N}\left(\boldsymbol{\beta}_{t,z_{d,n}}\right)$$



話者数の自動推定

- ☺ 自動的に話者数も推定できます
- 発話していない話者に対応するトピックの重み $z_{t,n,k}$ は学習と共に0に近づきます
- 従って"存在しない"話者に対応するトピック k' は以下を満たすかで判定できます

"存在する"話者に対応するトピック k

$$\frac{1}{K} \leq \sum_{t,n} z_{t,n,k}$$

"存在しない"話者に対応するトピック k'

$$\frac{1}{K} > \sum_{t,n} z_{t,n,k'}$$ (実際にはほぼ0になります)

隠れ変数とパラメータの推定

- 論文では変分ベイズ法(VB-EM)による解法が提案されています

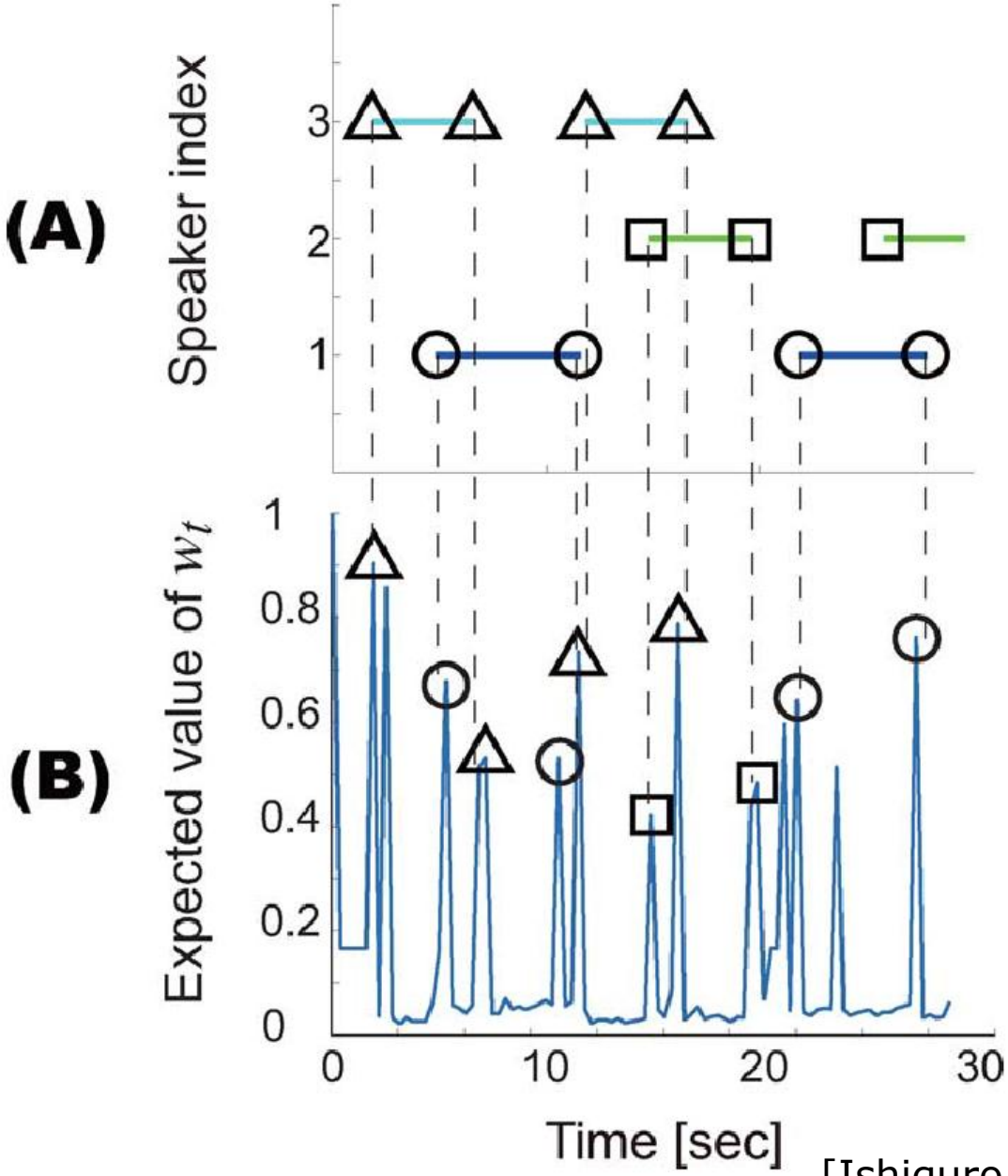
- 具体的な式は煩雑になるので省略します。必要な方は論文をチェックしてください

オンライン学習が自然に導かれます

- v_{tl} の定義から、θ_t (時刻tの話者分布)の学習には昔の分布の情報はほとんど影響しません
- すなわち、直近の情報だけを用いたオンライン(逐次)学習が可能となります

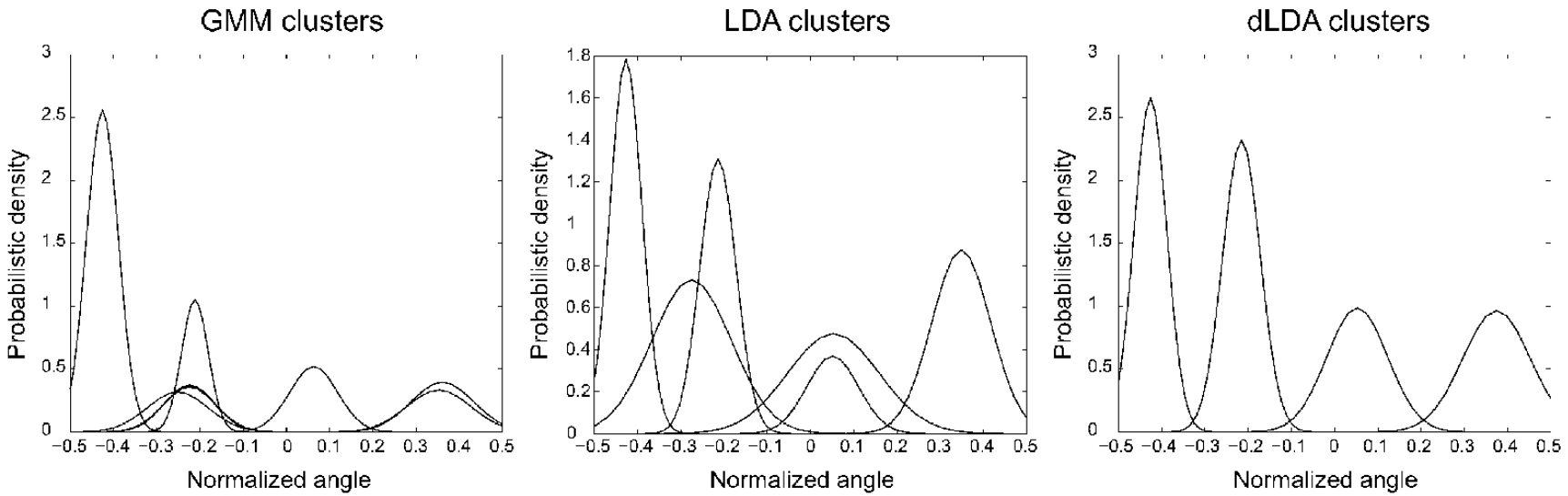
$$\boldsymbol{\theta}_t = (1 - w_t)\boldsymbol{\theta}_{t-1} + w_t\boldsymbol{\pi}_t$$

$$= \sum_{l=1}^{t} v_{tl}\boldsymbol{\pi}_l \qquad v_{tl} = w_l \prod_{m=l+1}^{t} (1 - w_m)$$



[Ishiguro, 2012]

話者4人のデータからのspeaker (topic)学習結果



[Ishiguro, 2012]

Dataset	[13]	[8]	GMM	LDA	dLDA(proposed)
CP1	21.9	(37.1)	55.8	32.7	**21.7**
*CP2	25.0	(35.8)	32.8	24.5	**19.7**
DC	**29.9**	(47.0)	60.6	48.0	31.0
CN	34.3	(56.4)	57.3	48.5	**34.1**
IS1000a	41.9	46.26	35.2	76.9	**32.2**
IS1001a	31.7	30.58	26.7	33.8	**23.7**
IS1001c	32.2	**12.07**	68.2	40.7	27.2
IS1006d	64.3	**54.56**	67.4	69.9	69.7
IS1008a	**13.1**	**5.13**	77.8	65.3	62.7
IS1008b	19.6	**16.47**	57.8	55.9	23.1
IS1008c	22.6	**12.09**	30.1	30.8	20.4
*IS1008d	**15.8**	20.83	21.9	32.1	**13.6**

[Ishiguro, 2012]

まとめ: Topic model for speaker diarizatoin

- トピックモデルにより、speaker diarization タスクを解決できます
- 簡単な時間発展モデルで話者の切り替わり(turn-taking)も自然にモデル化
- state-of-the-artの作りこんだモデルとcomparableの性能

その他の音声・音響データ応用

- Ohtsuka et al., “Bayesian Unification of Sound Source Localization and Separation with Permutation Resolution”, in Proc. AAAI, 2012.

- Yoshii and Goto, “A Nonparametric Bayesian Multiple Analyzer Based on Infinite Latent Harmonic Allocation”, IEEE Trans. ASLP, Vol. 20(3), pp. 717-730, 2012.

引用及び参考文献

- [Ishiguro, 2012] Ishiguro et al. , “Probabilistic Speaker Diarization with Bag-of-Words Representations of Speaker Angle Information”, IEEE Trans. ASLP, Vol. 20(2), pp. 447-460, 2012.
- [Araki, 2008] Araki et al., “A DOA based Speaker Diarization System for Real Meetings”, in Proc. Joint Workshop Hndns-Free Speech Comm. Microphone Arrays, 2008.
- [石黒 & 竹内, 2012] 石黒, 竹内, “特徴的な構造を抽出するデータマイニング技術”, NTT技術ジャーナル, Vol. 24, No. 9, 2012.