



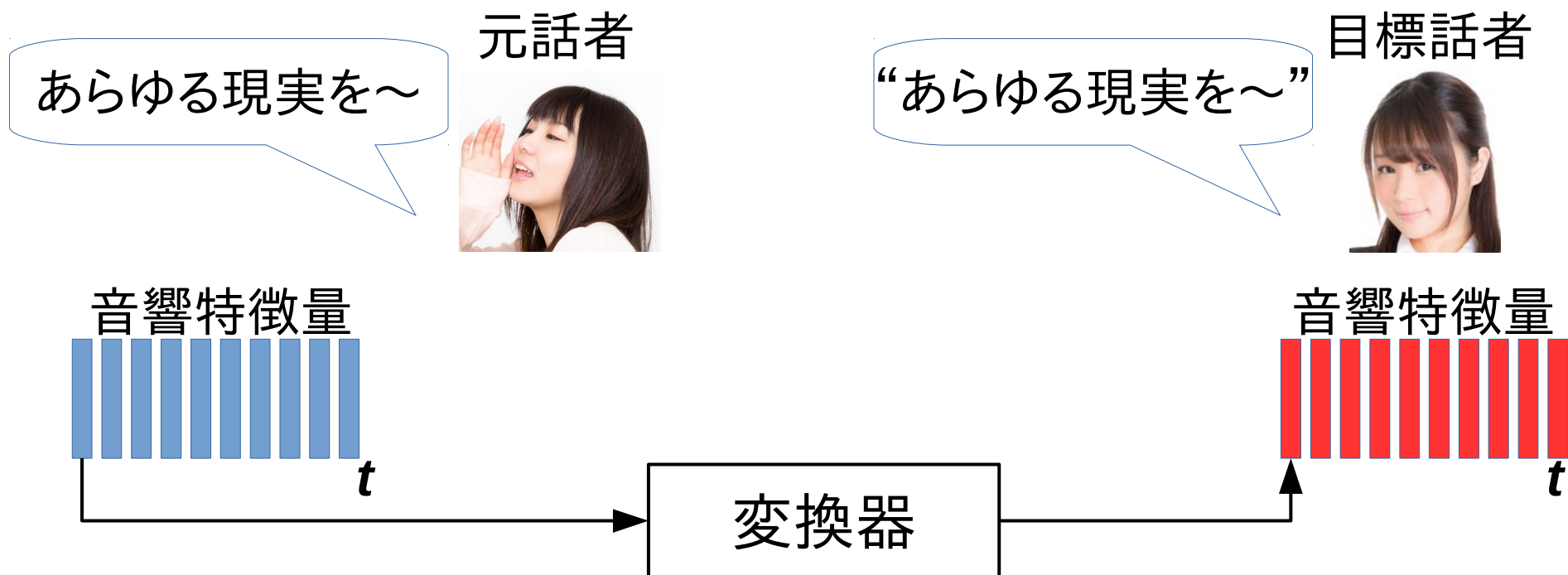
音声における「かわいさ」の制御 のための声質変換

日本大学大学院
博士前期課程 2 年
大野涼平



音声をかわいくしたい (1/2)

声質変換 [1, 2, 3]

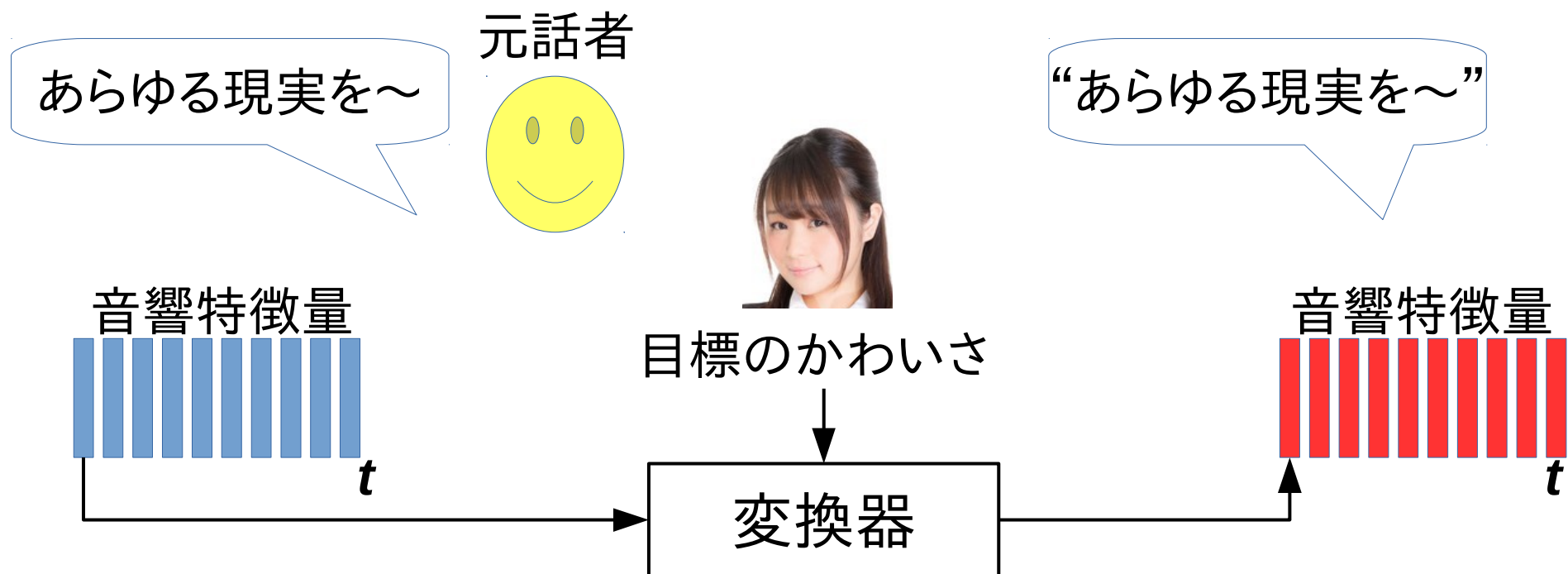


発話内容を保持したまま、
入力音声の声質を**目標話者**のものに変換する



音声をかわいくしたい (2/2)

本研究の目的



発話内容を保持したまま、
入力音声の声質を**目標のかわいさを持つ**ものに変換する



アウトライン

- 本研究の課題
- かわいさの評価と聴取時間の関係
- かわいさの主観評価値の予測
- 話者適応型RBMを用いたかわいさの制御
- まとめ



アウトライン

● 本研究の課題

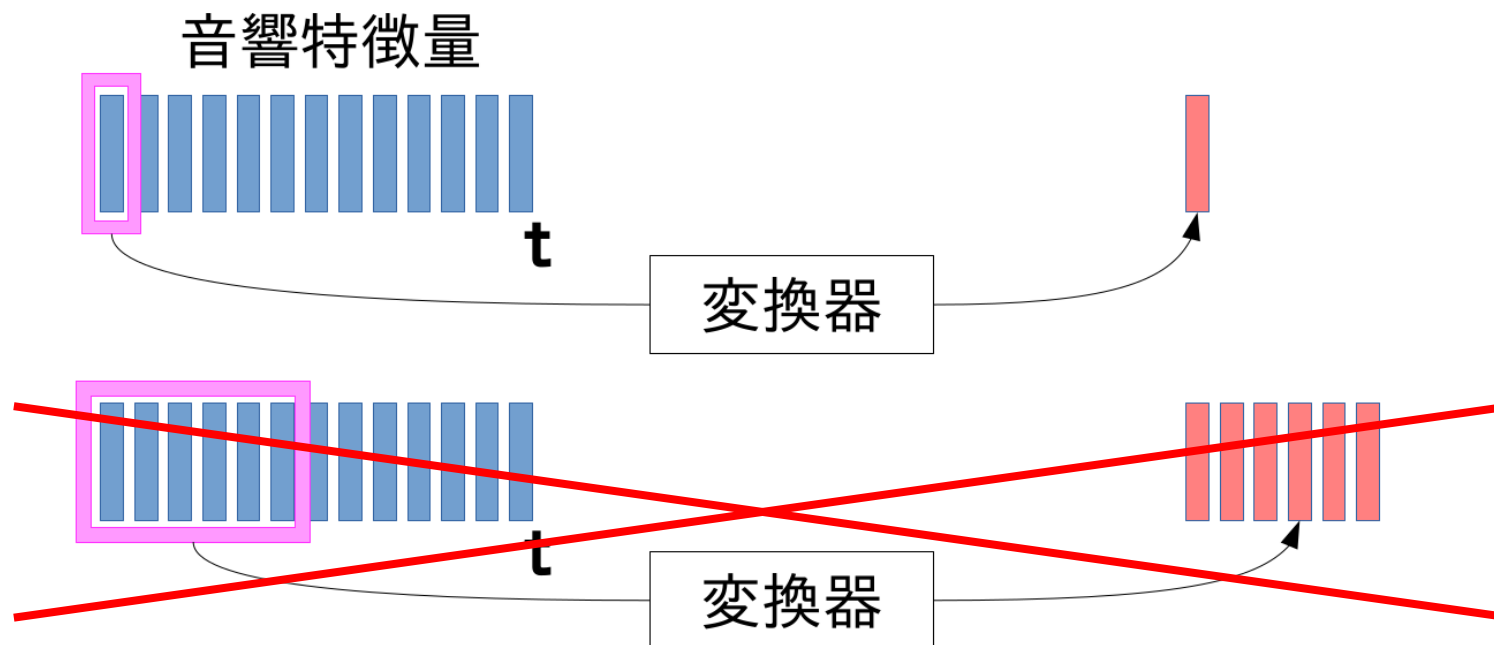
- かわいさの評価と聴取時間の関係
- かわいさの主観評価値の予測
- 話者適応型RBMを用いたかわいさの制御
- まとめ



局所的な特徴と「かわいさ」

研究課題

局所的な特徴の変換は「かわいさ」に寄与するか



つまり...

局所的な特徴は「かわいさ」を感じることに影響を与えるか

主観評価実験にて調査



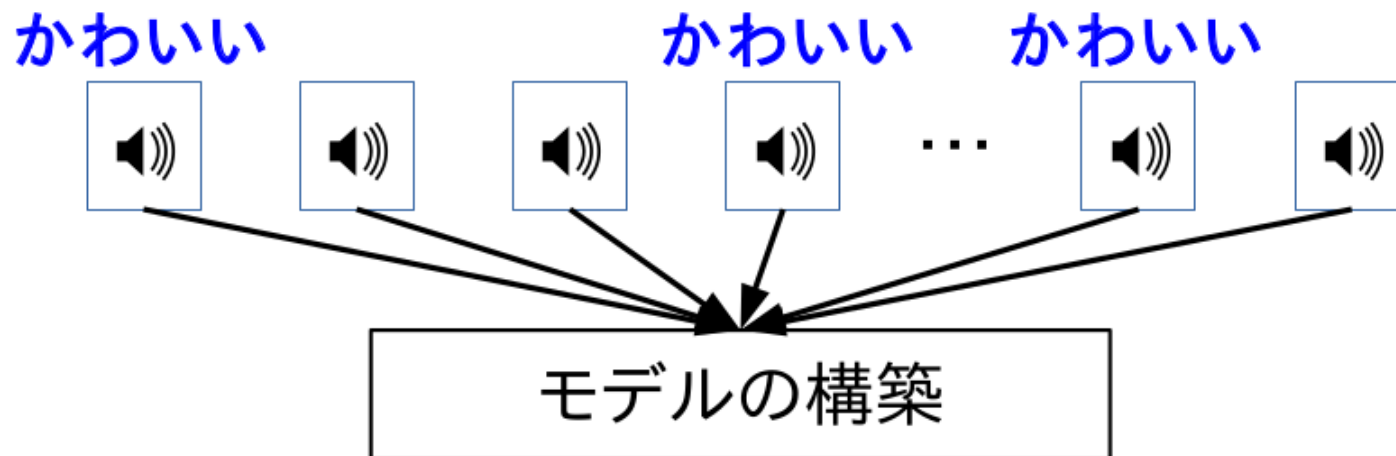
聴取者間の個人差への対応

研究課題

「どの音声からどれだけかわいさを感じるか」は聴取者次第

方策の1つ

ユーザの大量の話者に対する
主観評価値に基づいたモデルの構築



大量の主観評価に伴う時間的コストと疲労の対策として

協調フィルタリングによる主観評価値の予測

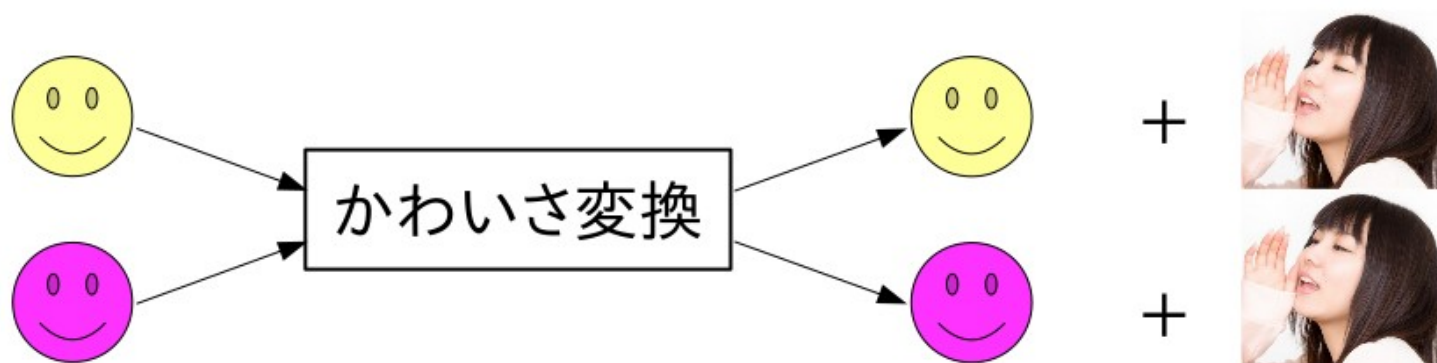
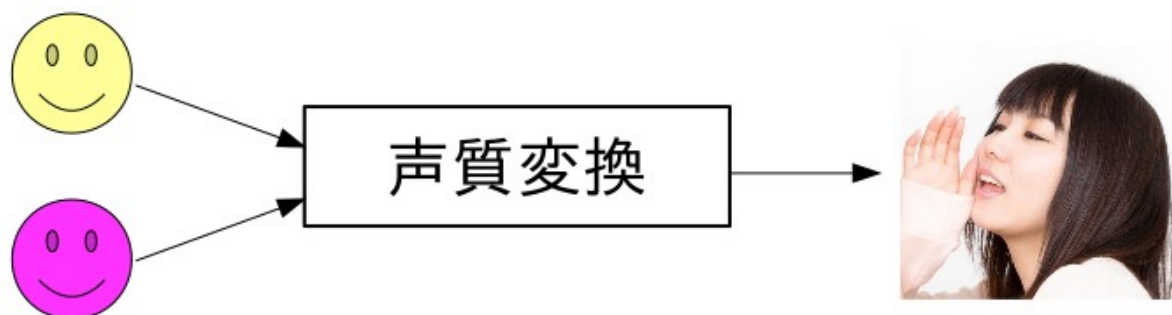


かわいさの個別の制御

研究課題

声質変換で生成した音声の話者性は一人

⇒ かわいさが起因する特徴だけを個別に制御したい



かわいさ非依存 かわいさ依存

話者適応型RBMで「かわいさ」を個別にモデル化



アウトライン

- 本研究の課題
- かわいさの評価と聴取時間の関係
- かわいさの主観評価値の予測
- 話者適応型RBMを用いたかわいさの制御
- まとめ



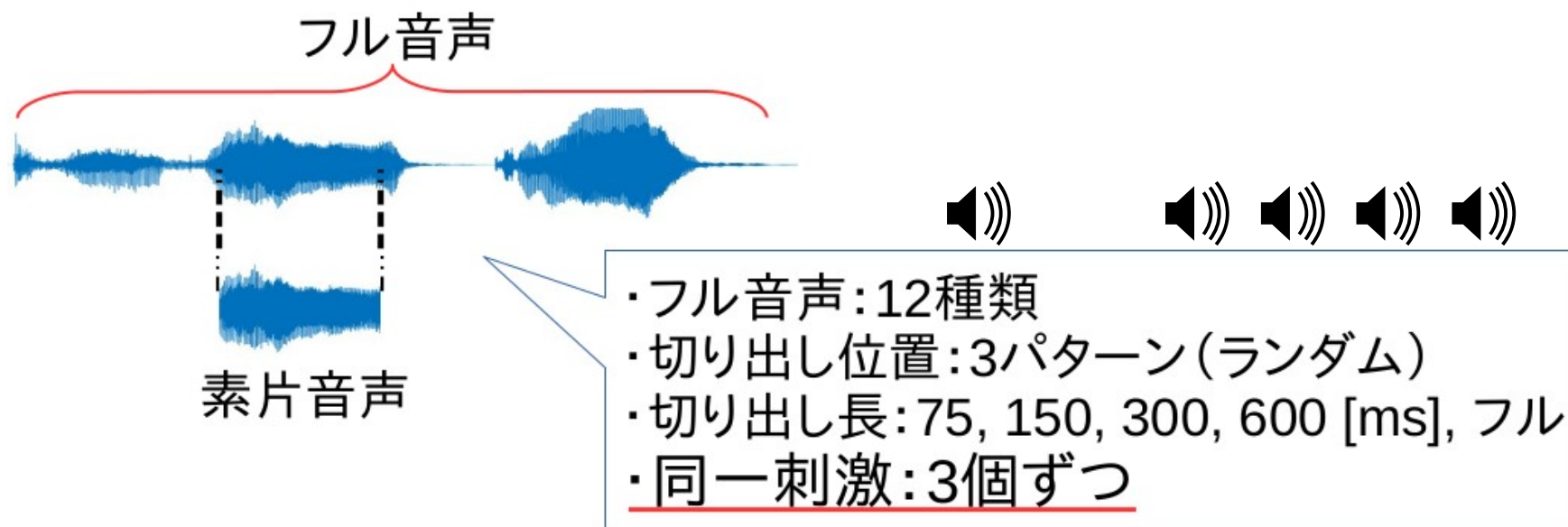
局所的な特徴の影響

評価と聴取時間

聴取時間が短くてもかわいさの評価が可能

⇒ 局所的な特徴の参照で評価が可能

様々な時間長の音声を用いた主観評価実験



3回の評価例(5段階)

1st	2nd	3rd
5	1	4

“安定性が低い”
(= 評価が難しい)



同一音声に対する評価

評価と聴取時間

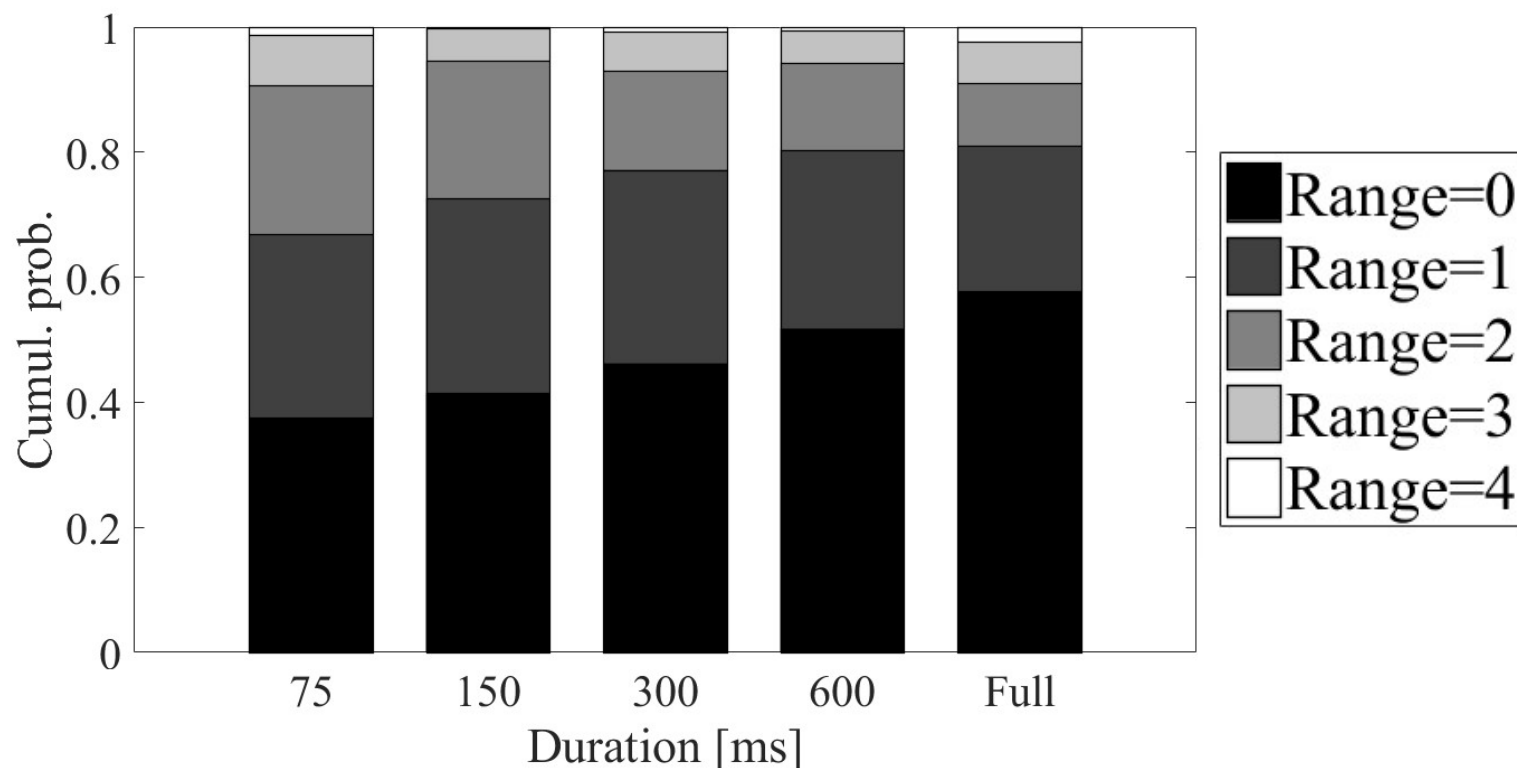


Fig. 同一音声に対する3回の評価のレンジ(累積度数)

- ・聴取時間が長い方が評価が安定する
- ・75 msでも66%以上は1以下に
⇒ 局所的な特徴量を参照している



高評価話者の素片音声への評価

評価と聴取時間

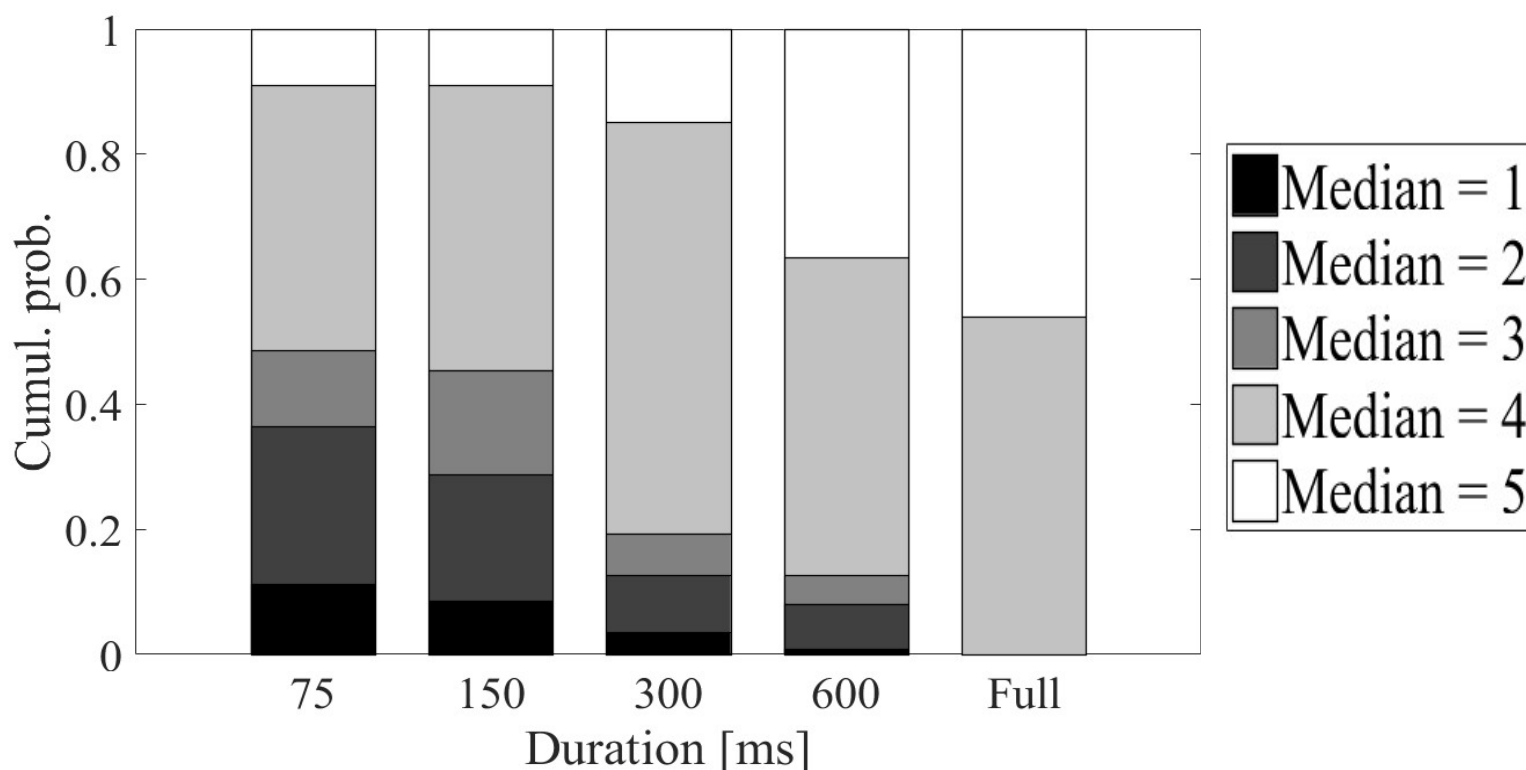


Fig. 高評価のフル音声の切り出し音声に対する3回の評価の中央値

- ・聴取時間が短くても「3」の割合は1割程度
 - ・75 msでも51%以上はフル音声の評価と一致
- ⇒ 局所的な特徴量を参照している



低評価話者の素片音声への評価

評価と聴取時間

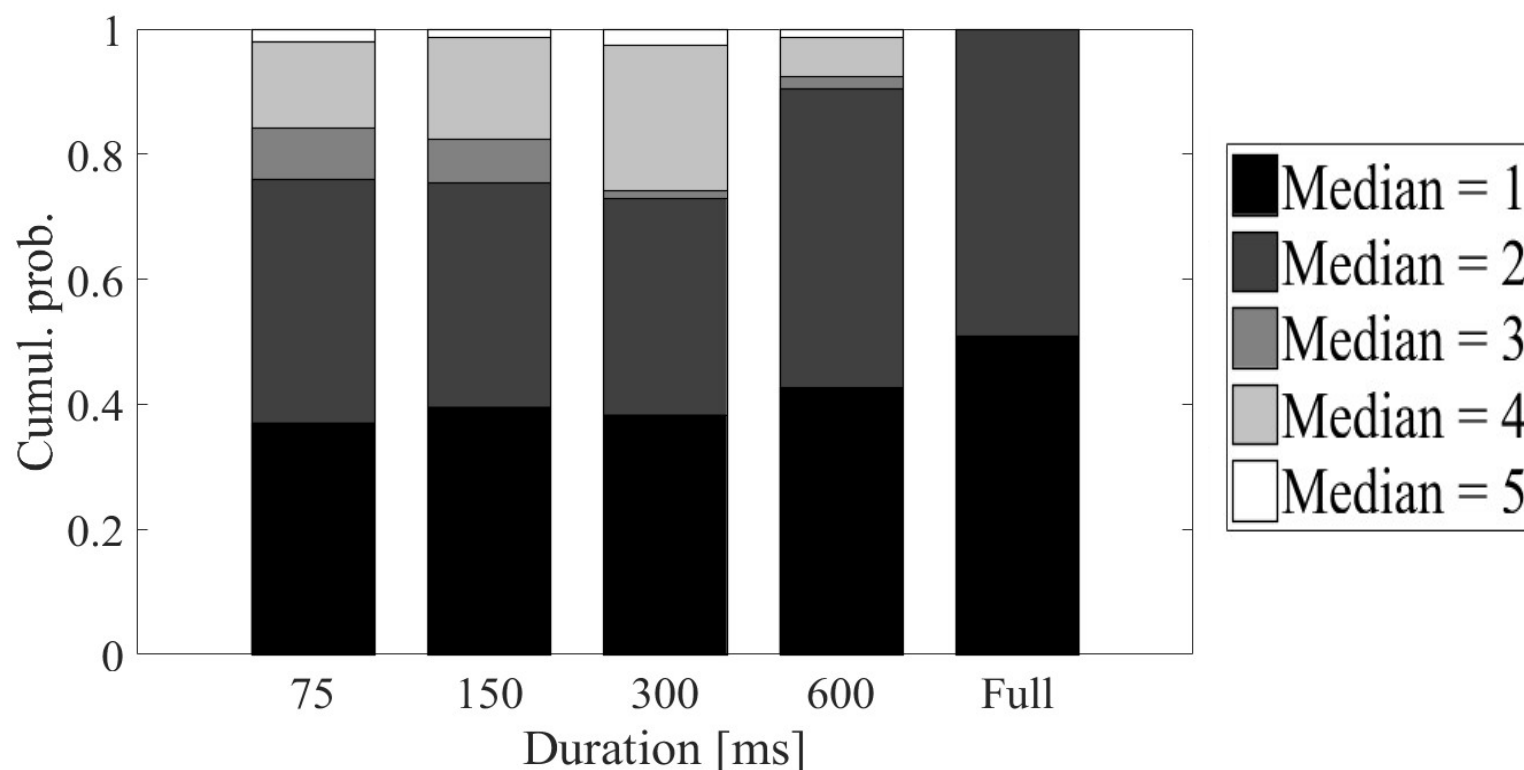


Fig. 低評価のフル音声の切り出し音声に対する3回の評価の中央値

- ・聴取時間が短くても「3」の割合は1割程度
 - ・75 msでも76%以上はフル音声の評価と一致
- ⇒ 局所的な特徴量を参照している



まとめ

評価と聴取時間

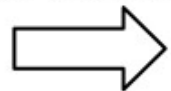
「かわいさ」の5段階評価で、同一刺激3回分の評価値の分布を分析

結果

75 msの音声刺激でも

- 評価のバラつきは66%以上が1以下
- 大半の評価がフル音声と一致

聴取時間が短くてもかわいさの評価が可能



局所的な特徴の参照で評価が可能

示唆された

局所的な特徴(スペクトル特徴量など)の変換が
かわいさの制御に貢献することが示唆された



アウトライン

- 本研究の課題
- かわいさの評価と聴取時間の関係
- かわいさの主観評価値の予測
- 話者適応型RBMを用いたかわいさの制御
- まとめ



主観評価の傾向

評価値の予測

再掲

「どの音声からどれだけかわいさを感じるか」は聴取者から

方策の1つ

ユーザの大量の話者に対する
主観評価値に基づいたモデルの構築

課題: 大量の主観評価に伴う時間的コストと疲労

聴取者間でかわいさの評価の傾向に類似性が存在？

⇒ 協調フィルタリング[4] による評価値の予測が可能

話者数70、評価者数42で、leave-one-out アプローチを用いて協調フィルタリングの精度を確認



協調フィルタリング

評価値の予測

予測実験

新規評価者の未評価値を, 他評価者らの評価値から予測する

	1	2	3	...	70
1	5	3	6	...	3
2	5	4	6	...	5
3	4	2	4	...	5
⋮	⋮	⋮	⋮	⋮	⋮
42	5	4	5	...	4

新規評価者の評価履歴

1	5			...	3
---	---	--	--	-----	---

60個をマスクし、予測対象とする ✕

類似している他評価者らの
評価履歴から予測する

2	5	4	6	...	5
---	---	---	---	-----	---

40	5	2	5	...	2
----	---	---	---	-----	---

✕ ランダムに60個を選択する



実験条件

評価値の予測

leave-one-out アプローチ

各評価者を新規評価者を選び、その評価者に対する
予測・推薦精度を計算する

Table 実験条件

類似度	相関係数
類似閾値	.75 ※1
予測精度	Mean Absolute Errors (MAEs), F-score, Precision, Recall
繰り返し	1000 回 ※2
備考	予測値が 5 以上の話者を推薦話者として出力し、 推薦話者に対して F-score, Precision, Recall を計算する。

※1 類似度がこれを超えた他評価者を類似評価者とする

※2 どの話者の評価を欠損させるかで推定結果が変わるため、
1000回分の試行結果の平均値を精度と見なす



全新規評価者の精度

評価値の予測

Table 精度の平均値 (標準偏差)

	MAEs	F-score	Precision	Recall
精度	.888 (.212)	.523 (.184)	.671 (.241)	.420 (.225)

- 予測精度: **MAEs < 0.9** (予測誤差が0.9未満)
- 推薦精度: $0.4 < \text{F-score} < 0.6$



新規評価者毎の精度

評価値の予測

(a) Top 25%

User's ID	11	14	19	20	22	24	30	31	37	40
Mean of #similar	16.2	13.6	10.3	9.1	15.6	17.2	11.5	15.3	16.7	14.8
MAE	.629	.708	.812	.704	.690	.766	.766	.725	1.092	.882
F-value	.745	.751	.748	.800	.736	.740	.773	.713	.755	.720
Precision	.862	.893	.826	.846	.862	.964	.889	.877	.909	.831
Recall	.702	.637	.690	.760	.684	.643	.698	.662	.698	.680

(b) Bottom 25%

User's ID	10	12	18	21	25	29	32	33	36	38
Mean of #similar	.2	5.2	14.0	2.2	.7	10.0	3.3	4.6	5.8	1.5
MAE	.961	.914	.680	.818	1.015	.878	1.309	1.023	.914	1.860
F-value	.170	.341	.338	.324	.190	.345	.307	.284	.318	.245
Precision	.182	.373	.605	.265	.134	.572	.567	.317	.435	.354
Recall	.015	.267	.260	.220	.046	.243	.127	.195	.225	.062



新規評価者毎の精度

評価値の予測

(a) Top 25%

User's ID	11	14	19	20	22	24	30	31	37	40
Mean of #similar	16.2	13.6	10.3	9.1	15.6	17.2	11.5	15.3	16.7	14.8
MAE	.629	.708	.818	.802	.768	.768	.785	.808	.882	.882
F-value	.745	.893	.820	.840	.802	.904	.889	.877	.909	.831
Precision	.862	.893	.820	.840	.802	.904	.889	.877	.909	.831
Recall	.702	.637	.690	.760	.684	.643	.698	.662	.698	.680

類似評価者数が10人以上 (ID 20以外)

(b) Bottom 25%

User's ID	10	12	18	21	25	29	32	33	36	38
Mean of #similar	.2	5.2	14.0	2.2	.7	10.0	3.3	4.6	5.8	1.5
MAE	.961	.914	.688	.878	.878	.878	.878	.878	.878	.860
F-value	.170	.170	.170	.170	.170	.170	.170	.170	.170	.170
Precision	.182	.182	.182	.182	.182	.182	.182	.182	.182	.182
Recall	.015	.267	.200	.220	.040	.243	.127	.193	.223	.062

類似評価者数が1桁 (ID 18, 29以外)
ID 10, 25は1人未満



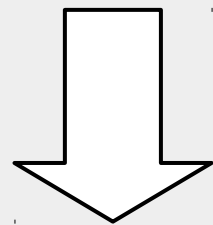
予測・推薦精度と評価の傾向

評価値の予測

- 推薦精度が**高い** / **低い**新規評価者は
 - 事前の主観評価:**50%以上** / **30%未満**の話者に5以上を付けた
 - 類似評価者数:**10人以上**(ID 20以外) / **1桁**(ID 18, 29以外)

主観評価で高評価を付ける割合 と 類似評価者数

⇒ 正の相関 (0.65)



類似評価者数が少ない場合、
協調フィルタリングは機能しにくい

「かわいい」と感じる話者が少ない場合、評価の予測が困難



まとめ

評価値の予測

話者数70、評価者数42で、leave-one-out アプローチを用いて協調フィルタリングの精度を確認

結果

- 予測精度 (MAEs) は0.9未満
- 話者推薦におけるPrecisionは0.67
- 高評価を与えた話者の数と類似評価者数に正の相関(0.65)

聴取者間でかわいさの評価の傾向に類似性が存在：

示唆された

⇒ 協調フィルタリングによる評価値の予測が可能

ユーザの大量の話者に対する主観評価値に基づいた
分析に貢献できると期待される



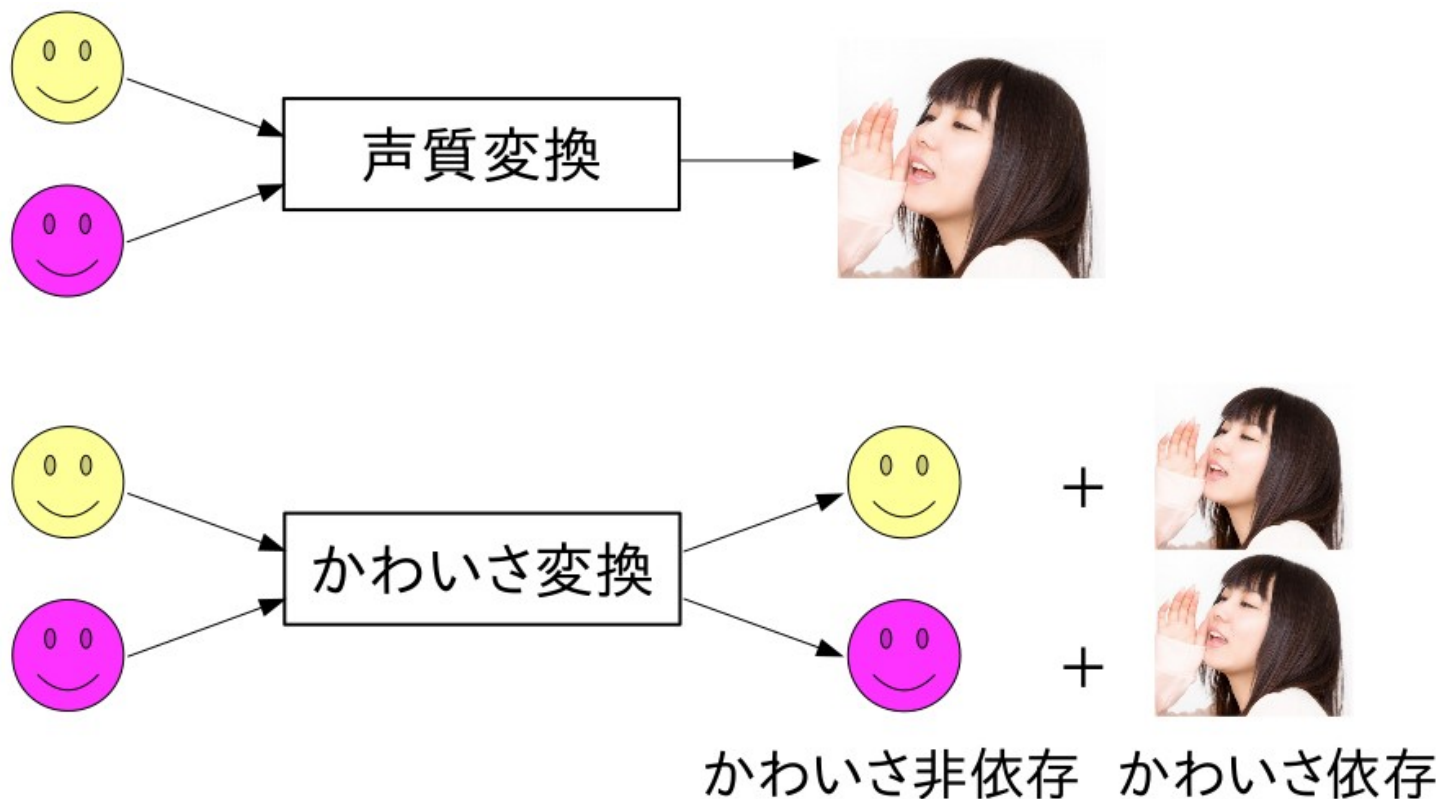
アウトライン

- 本研究の課題
- かわいさの評価と聴取時間の関係
- かわいさの主観評価値の予測
- 話者適応型RBMを用いたかわいさの制御
- まとめ



かわいさの個別制御

ARBM



声質変換で生成した音声の話者性は一人

⇒ かわいさが起因する特徴だけを個別に制御したい

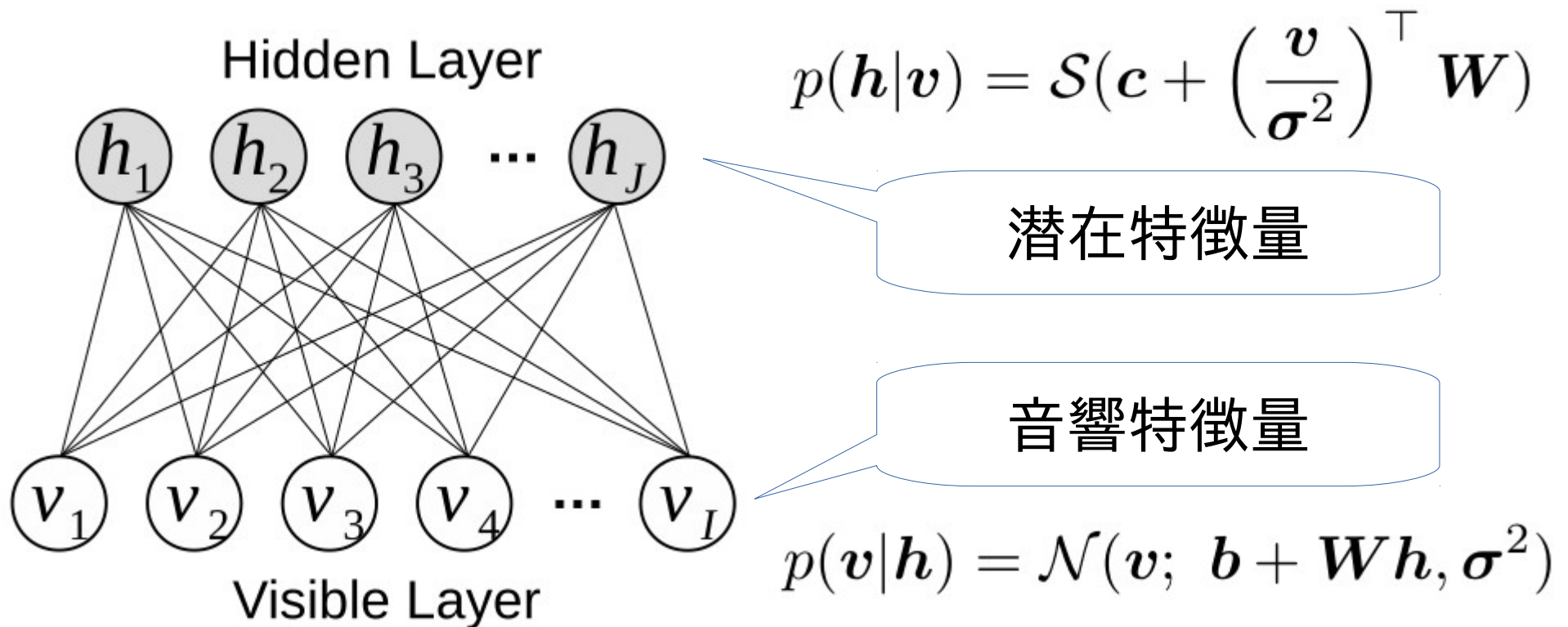
「かわいさ」の度合いによるクラスタ依存のパラメータを含める学習法を提案



Restricted Boltzmann Machine

ARBM

Restricted Boltzmann Machine [6, 7, 8]



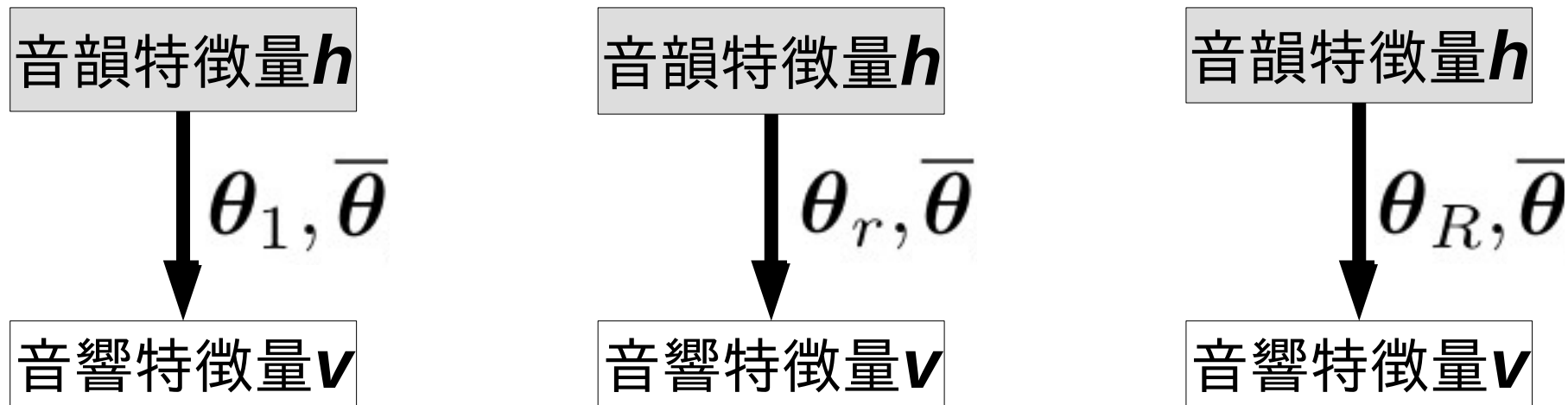
可視層の尤度を最大化するように学習する



話者適応型RBMと提案法 2/2

ARBM

$\bar{\theta}$... 話者非依存パラメータ
 θ_r ... 話者依存パラメータ



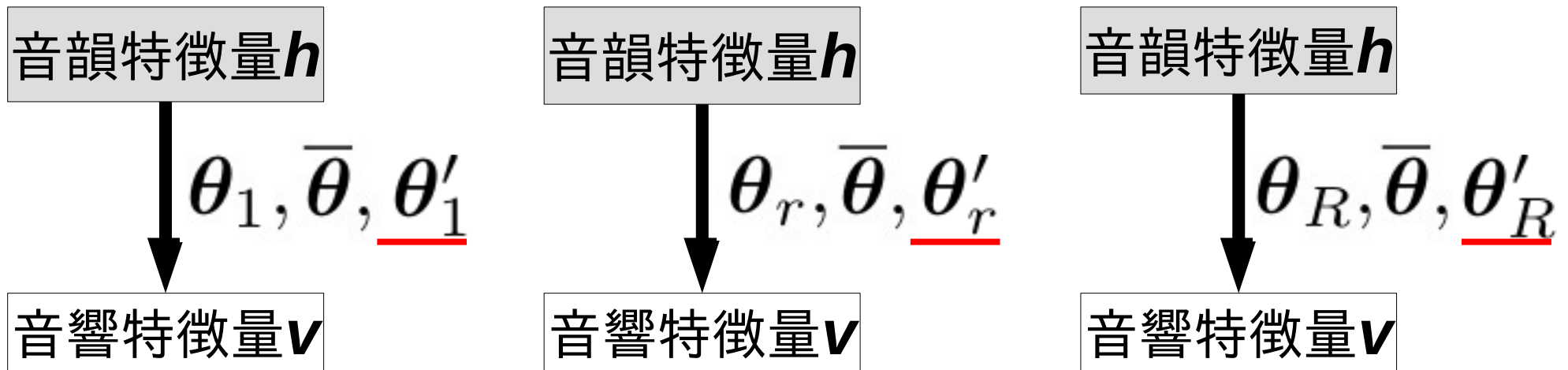
- 音韻特徴量 (隠れ層) が話者依存・非依存パラメータを用いたある過程を経て、音響特徴量 (可視層) が生成される
- 音韻特徴量は可視層からフィードフォワード処理で推定
- 話者依存パラメータを適切に差し替える事で**所望の話者の音声**の生成が可能



話者適応型RBMと提案法 2/2

ARBM

$\theta'_r \dots$ クラス依存パラメータ
($\theta'_r = \{\theta_{K'}, \theta_{\bar{K}'}\}$)



- かわいい話者、かわいくない話者の2クラスに依存するパラメータを生成過程で考慮する
- 各話者はかわいい話者なのかかわいくない話者なのかユーザによって定められる



実験条件

ARBM

スペクトル特徴量	24次のメルケプストラム
分析合成	WORLD (DC4 edition) [9, 10]
学習エポック	事前学習:100回 事後学習a:100回 事後学習b:50回
学習法	確率的勾配法 ミニバッチ:100R 学習率:0.01 モーメンタム:0.9
モデル	可視層:24ユニット 隠れ層:8ユニット
クラス	クラスK:主観評価値が6以上の話者 クラス $\bar{K}$:主観評価値が2以下の話者



主観評価実験

ARBM

クラス $\bar{K}$ の話者  を次のように変換する

手法	
X2Y 	クラス $\bar{K}$ の話者へ変換する (ARBMと等価)。
X2X' 	入力話者とは異なるクラス $\bar{K}$ の話者へ変換する (ARBMと等価)。
PRO 	入力話者とは異なるクラス $\bar{K}$ の話者へ変換する。 ただし、クラス依存パラメータのみをクラス $\bar{K}$ の ものに差し替える。

※ **X2X'**と**PRO**の出力話者は同一話者とする

以下を確かめる

- **PRO**の生成音声が他手法よりかわいい
- **PRO**の生成音声に変換前よりかわいい



主観評価結果1

ARBM

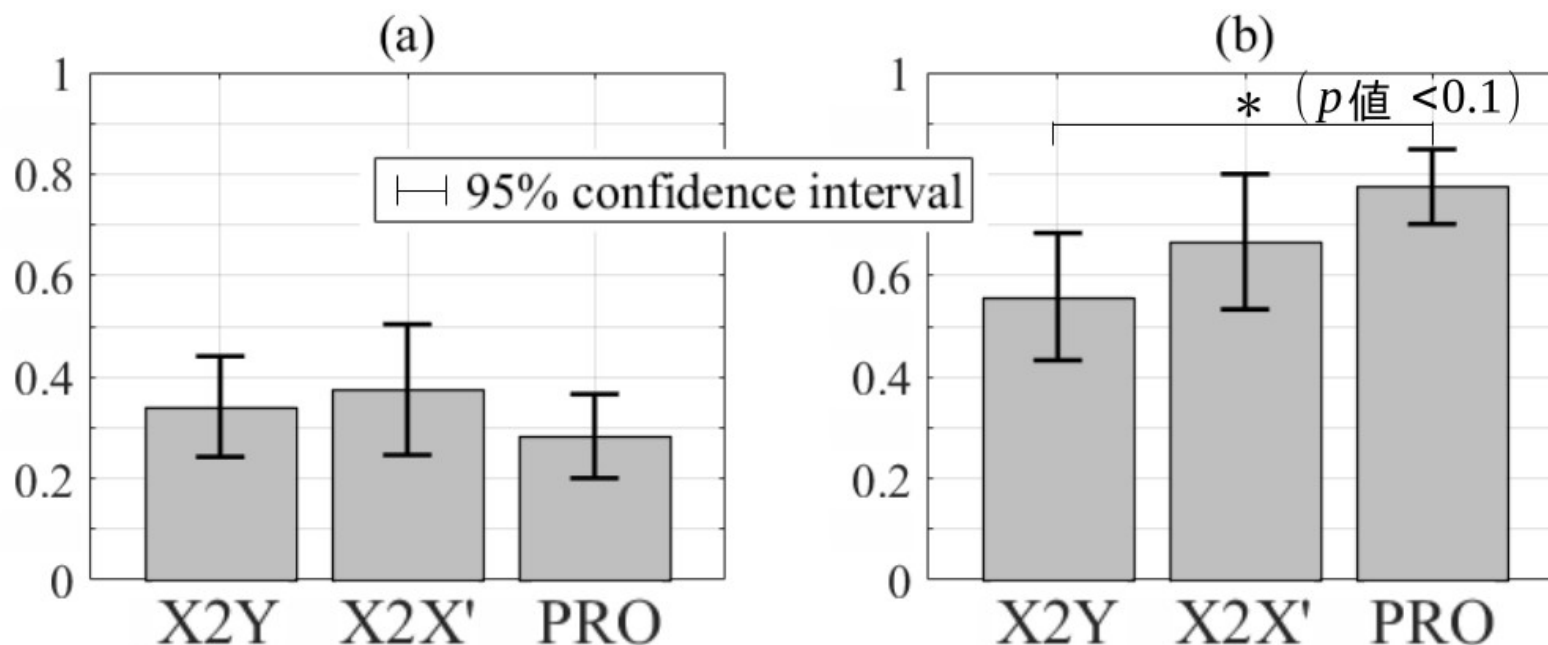


Fig. 順位付けにおいて、(a) 1位に選択された割合 (b) 2位以内に選択された割合

- **PRO**による生成音声の80%以上が2位以内
- かわいさは平均で3.1（7段階:1~7）
- 60%以上が3以上の評価



主観評価結果2

ARBM

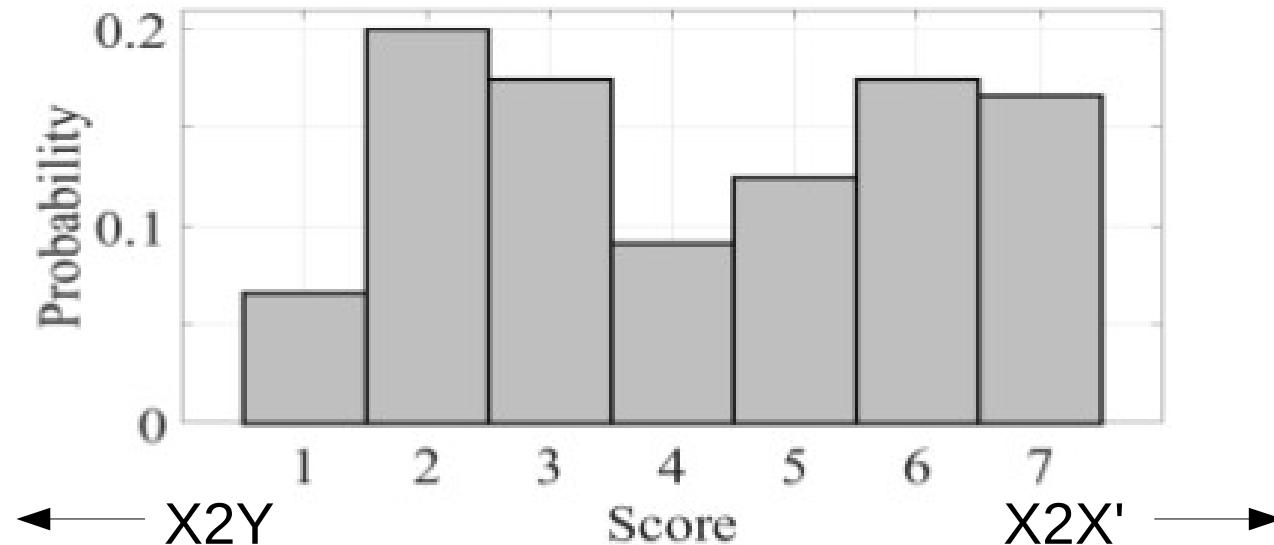


Fig. 話者性に関するスコアの分布

X2X'と同じ目標話者に変換しているにも関わらず、**PRO**の音声の話者性は44.2%が**X2Yのものに近い**

⇒ かわいさの特徴を制御したことで話者性が変化した



まとめ

ARBM

「かわいさ」の度合いによるクラスタ依存のパラメータを含める学習法を提案

結果

かわいくない話者に変換しているが、かわいさを制御したことで

- 従来手法よりも80%近くの回答で**かわいさの順位が上位**
- 主観評価値が平均で**3.1** (60%以上が3以上)
- 話者性がかわいいクラスの話者に近いというのが44.2%

以下を確かめる

- **PRO**の生成音声が他手法よりかわいい
- **PRO**の生成音声に変換前よりかわいい
- **PRO**の生成音声の話者性がクラスKの話者に近い

示唆された



アウトライン

- 本研究の課題
- かわいさの評価と聴取時間の関係
- かわいさの主観評価値の予測
- 話者適応型RBMを用いたかわいさの制御

● まとめ



音声をかわいくしたい



- ☆ 「かわいさ」は局所的な特徴量でも表現可能か
75msの音声でもかわいさの評価が可能

⇒ 局所的な特徴量の制御でかわいさの制御が可能

示唆された

- ☆ ユーザはどの声を「かわいい」と感じるか

協調フィルタリングによる「かわいい声」の推定は可能

示唆された

- ☆ 話者適応型声質変換で「かわいさ」の適応
従来手法よりもかわいさが優位



引用

1. T. Toda *et al.*, “**Voice conversion based on maximum likelihood estimation of spectral parameter trajectory**,” *IEEE Trans. on ASLP*, 15(8), pp. 2222–2235, 2007.
2. Y. Saito, *et al.*, “**Voice Conversion Using Input-to-Output Highway Networks**,” *IEICE Trans. on Information and Systems*, E100-D(8), pp. 1925–1928, 2017.
3. C. C. Hsu, *et al.*, “**Voice Conversion from Unaligned Corpora Using Variational Autoencoding Wasserstein Generative Adversarial Networks**,” *Proc of INTERSPEECH*, pp. 3364–3368, 2017.
4. P. Resnick, *et al.*, “**GroupLens: an open architecture for collaborative filtering of netnews**,” *In Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pp. 175–186, 1994.
5. T. Nakashika *et al.*, “**Non-Parallel Training in Voice Conversion Using an Adaptive Restricted Boltzmann Machine**,” *IEEE Trans. on ASLP*, 24(11), pp. 2032--2045, 2016.
6. K. Cho *et al.*, “**Improved learning of Gaussian-Bernoulli restricted Boltzmann machines**,” in *Artificial Neural Networks and Machine Learning – ICANN 2011*. 2011, pp. 10--17, Springer.



引用

7. G. E. Hinton, *et al.*, “**A fast learning algorithm for deep belief nets,**” *Neural computation*, 18(7), pp. 1527–1554, 2006.
8. G. E. Hinton, “**A practical guide to training restricted Boltzmann machines,**” *Tech. Rep. UTML TR 2010-003*, 2010.
9. M. Morise, *et al.*, “**WORLD: A vocoder-based high-quality speech synthesis system for real-time applications,**” *IEICE Trans. on Information and Systems*, 99(7), pp. 1877–1884, 2016.
10. M. Morise “**D4C, a band-aperiodicity estimator for high-quality speech synthesis,**” *Speech Communication*, 84, pp. 57–65, 2016.



修士課程業績一覧 (1)

研究会

1. 大野涼平, 森勢将雅, 北原鉄朗: “音声における「かわいらしさ」の知覚と聴取時間の関係性の検討,” 情報処理学会音楽情報科学研究会, vol. 2016-MUS-111, no. 50, pp. 1–5, 2016. (学生奨励賞)

全国大会 (主著)

1. 大野涼平, 高道慎之介, 森勢将雅, 北原鉄朗: “話者適応型 RBM を用いたユーザが所望するかわいい音声への声質変換,” 日本音響学会 2018 年春季研究発表会, 2018. (発表予定)
2. 大野涼平, 高道慎之介, 森勢将雅, 北原鉄朗: “音声の「かわいさ」における主観的傾向の一分析,” 日本音響学会 2017 年秋季研究発表会, pp. 401–402, 2017.
3. 大野涼平, 高道慎之介, 森勢将雅, 北原鉄朗: “統計的声質変換における印象評価の調査,” 日本音響学会 2017 年春季研究発表会, pp. 371–372, 2017.



修士課程業績一覽 (2)

国際会議

1. Ohno, R., Morise, M., and Kitahara, T.: “Relationship between perception of cuteness in female voices and their durations,” *Lecture Notes in Computer Science (in Proc. of SPECOM)*, LNCS, 10458, pp. 642–650, 2017.
2. Ohno, R., Morise, M., and Kitahara, T.: “The relationship between perception of cuteness and duration of voices,” *Joint Meeting : Acoustical Society of America and Acoustical Society of Japan.*, vol. 140, no. 4, p. 3399, 2016.



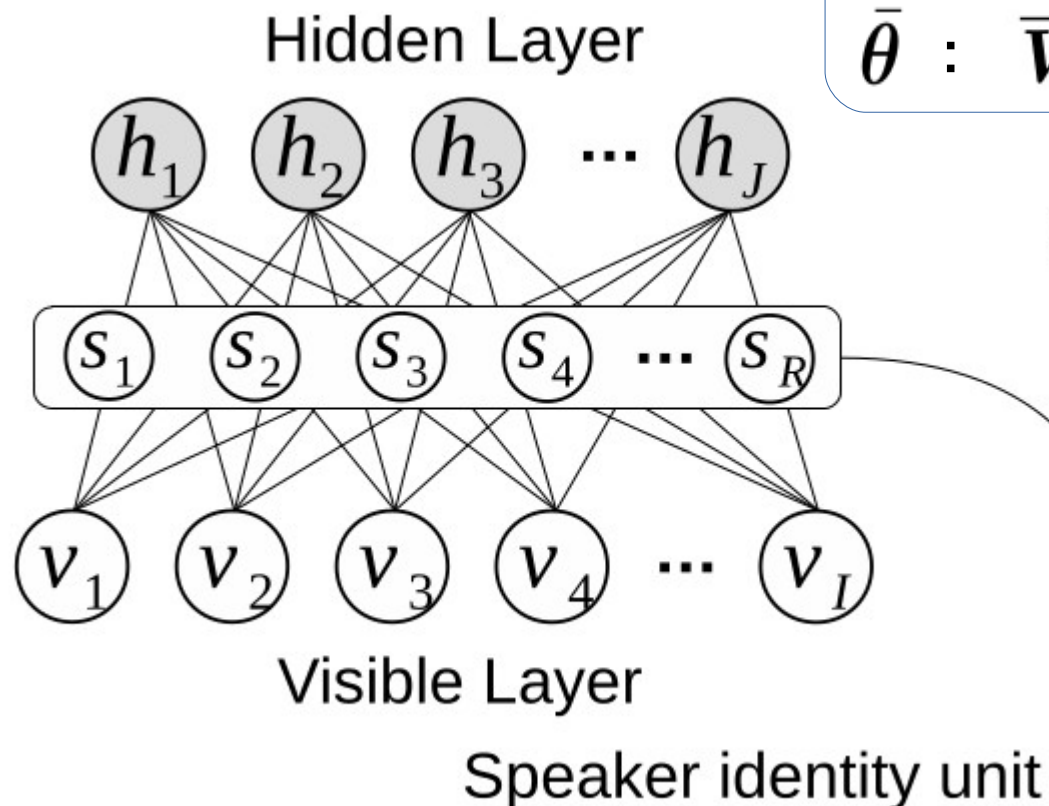
付録



話者適応型RBM

ARBM

話者適応型RBM [5]



$\theta_r : \mathbf{A}_r, \mathbf{b}_r, \mathbf{c}_r$ 話者依存パラメータ

$\bar{\theta} : \bar{\mathbf{W}}, \bar{\mathbf{b}}, \bar{\mathbf{c}}$ 話者非依存パラメータ

$$p(\mathbf{h}|\mathbf{v}) = \mathcal{S}(\mathbf{c} + \left(\frac{\mathbf{v}}{\sigma^2}\right)^\top \mathbf{W})$$

$$\mathbf{W} = \mathbf{A}_r \bar{\mathbf{W}}$$

$$\mathbf{b} = \mathbf{b}_r + \bar{\mathbf{b}}$$

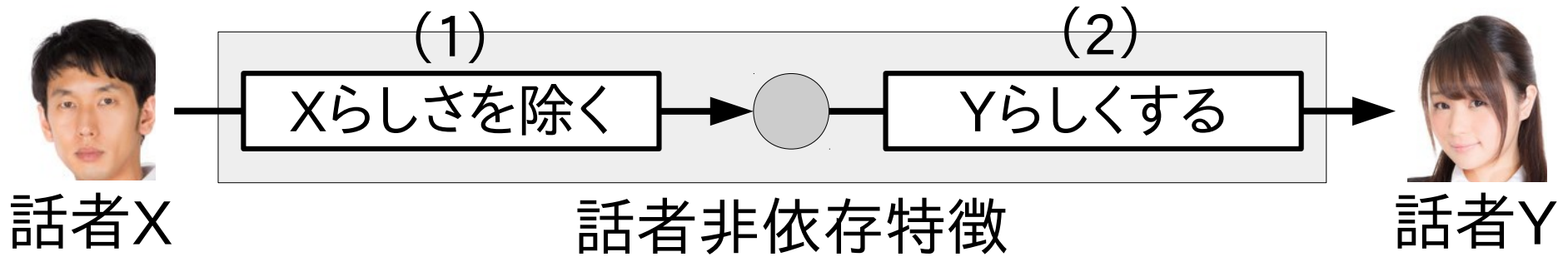
$$\mathbf{c} = \mathbf{c}_r + \bar{\mathbf{c}}$$

$$p(\mathbf{v}|\mathbf{h}) = \mathcal{N}(\mathbf{v}; \mathbf{b} + \mathbf{W}\mathbf{h}, \sigma^2)$$



音声変換への応用 (1)

ARBM



$$1) \hat{\mathbf{h}}^{(t)} \triangleq \operatorname{argmax}_{\mathbf{h}^{(t)}} p(\mathbf{h}^{(t)} | \mathbf{v}_x^{(t)})$$

$$\mathbf{W} = \mathbf{A}_x \overline{\mathbf{W}}$$

$$\mathbf{c} = \mathbf{c}_x + \overline{\mathbf{c}}$$

$$2) \hat{\mathbf{v}}_y^{(t)} \triangleq \operatorname{argmax}_{\mathbf{v}_y^{(t)}} p(\mathbf{v}_y^{(t)} | \hat{\mathbf{h}}^{(t)})$$

$$\mathbf{W} = \mathbf{A}_y \overline{\mathbf{W}}$$

$$\mathbf{b} = \mathbf{b}_y + \overline{\mathbf{b}}$$

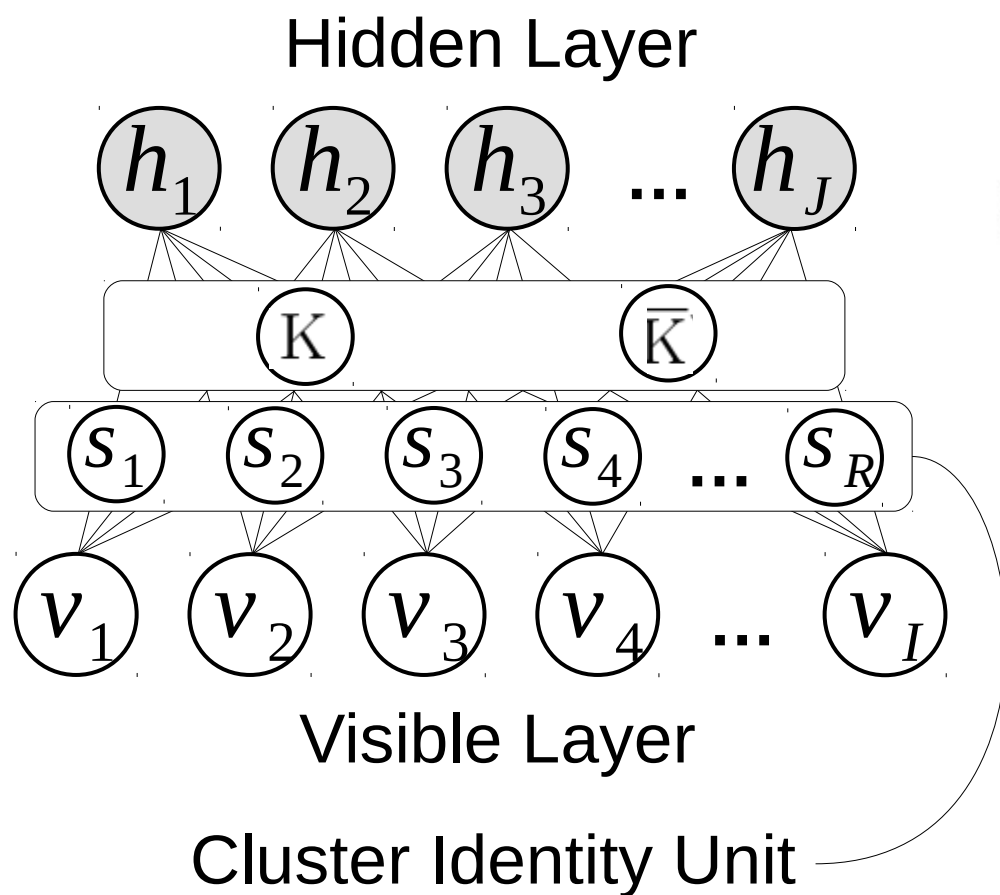


かわいさ適応型RBM

ARBM

話者適応型RBM [5]

$\theta_k : \mathbf{H}_k, \mathbf{b}'_k, \mathbf{c}'_k$ クラス依存パラメータ



$$p(\mathbf{h}|\mathbf{v}) = \mathcal{S}(\mathbf{c} + \left(\frac{\mathbf{v}}{\sigma^2}\right)^\top \mathbf{W})$$

$$\mathbf{W} = \mathbf{H}_k \mathbf{A}_r \overline{\mathbf{W}}$$

$$\mathbf{b} = \mathbf{b}'_k + \mathbf{b}_r + \overline{\mathbf{b}}$$

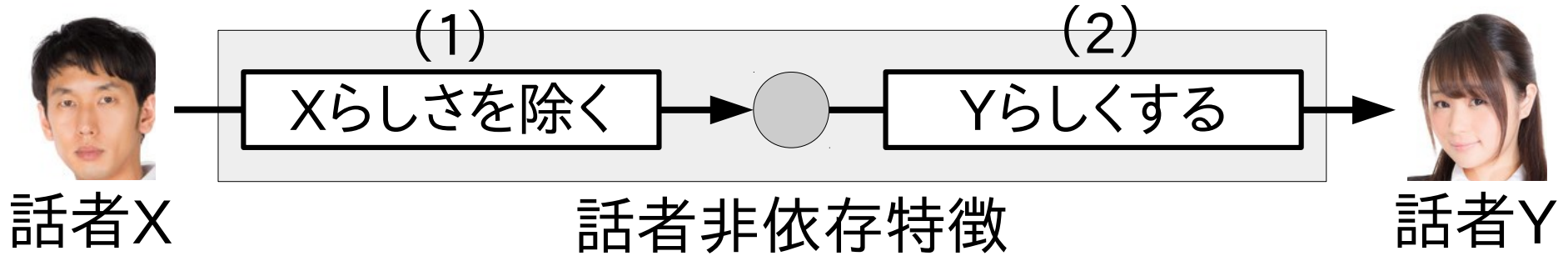
$$\mathbf{c} = \mathbf{c}'_k + \mathbf{c}_r + \overline{\mathbf{c}}$$

$$p(\mathbf{v}|\mathbf{h}) = \mathcal{N}(\mathbf{v}; \mathbf{b} + \mathbf{W}\mathbf{h}, \sigma^2)$$



音声変換への応用 (2)

ARBM



$$1) \hat{h}^{(t)} \triangleq \operatorname{argmax}_{h^{(t)}} p(h^{(t)} | v_x^{(t)})$$

$$W = H_x A_x \bar{W}$$

$$b = b'_x + b_x + \bar{b}$$

$$2) \hat{v}_y^{(t)} \triangleq \operatorname{argmax}_{v_y^{(t)}} p(v_y^{(t)} | \hat{h}^{(t)})$$

$$W = H_y A_y \bar{W}$$

$$b = b'_y + b_y + \bar{b}$$



Vocoder

音声



分解

基本周波数

スペクトル包絡

非周期性指標

基本周波数

スペクトル包絡

非周期性指標

再合成

変換音声

