おしゃべりなコンピュータ～音声合成技術の現在と未来～

山岸順一 准教授

国立情報学研究所 コンテンツ科学研究系

英国エジンバラ大学 音声技術研究所

本日の講義の構成

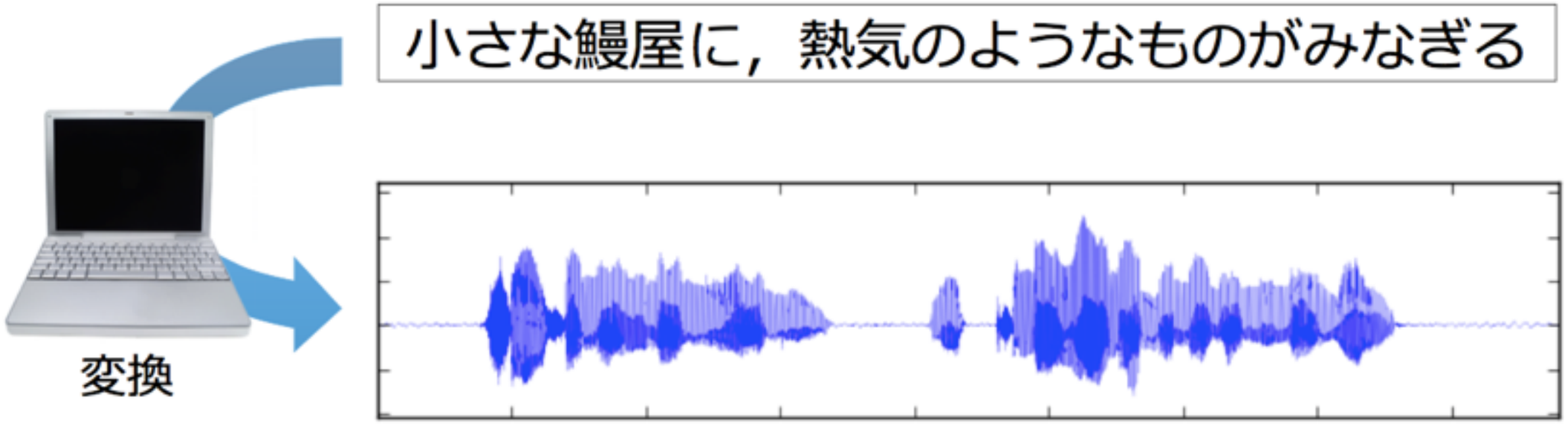
- 音声合成
  - 音声合成：音声と文章の対応付け
    - 文章の言語情報抽出
      - 音素、国際音声記号、レキシコン等
    - 音声の音響特徴量抽出
      - スペクトル、フォルマント、音素との対応、スペクトル包絡
    - 言語情報と音響特徴量の対応付け
      - 様々な音声合成方式
      - 統計的音声合成方式
- 統計的音声合成の種々の応用

現在の音声技術

- 音声情報処理技術はだいぶ普及してきました
    - 音声認識
    - 音声検索
        - Google voice search
    - 音声翻訳
        - Google translation
- 音声対話エージェント
    - Siri, しゃべってコンシェルジュ
- 音声合成
    - ボーカロイド

音声合成

- テキスト音声合成：入力テキストを自然で聞き取りやすい音声へ変換



- なぜ必要か？
  - 情報パネル上の表示を見ることができない（カーナビゲーション）
  - 情報パネルが小さい （携帯端末や時計端末の音声対話エージェント）
  - 視覚・音声の障害 （スクリーンリーダ、会話補助システム）

音声合成：文章と音声の対応付け（マッピング）

小さな鰻屋に，熱気のようなものがみなぎる

↑

文章の言語的情報抽出

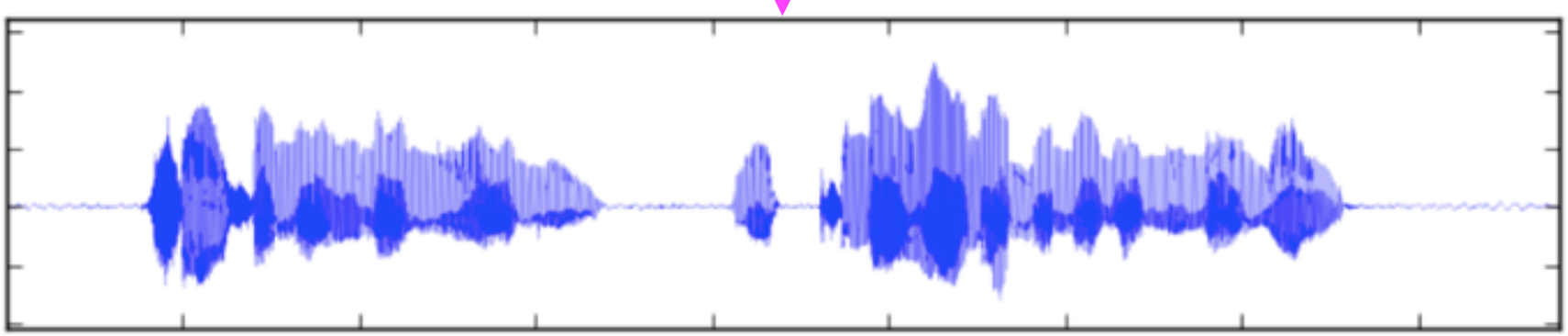
↑

言語情報と音響的特徴量の対応付け

↑

音声の音響的特徴量抽出

↓



音声合成：文章と音声の対応付け（マッピング）

小さな鰻屋に，熱気のようなものがみなぎる

文章の言語的情報抽出

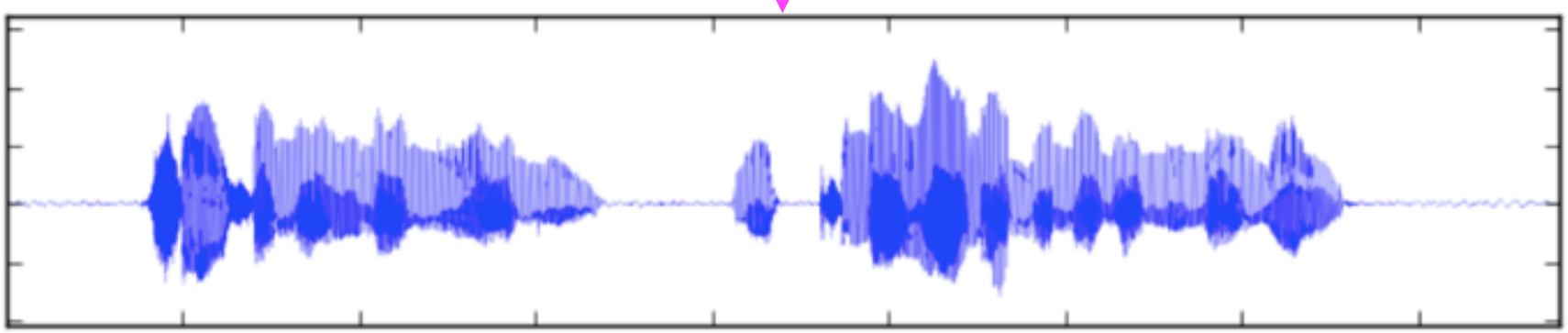
言語学・
自然言語処理

言語情報と音響的特徴量の対応付け

機械学習

音声の音響的特徴量抽出

信号処理



音声合成：文章と音声の対応付け（マッピング）

小さな鰻屋に，熱気のようなものがみなぎる

文章の言語的情報抽出

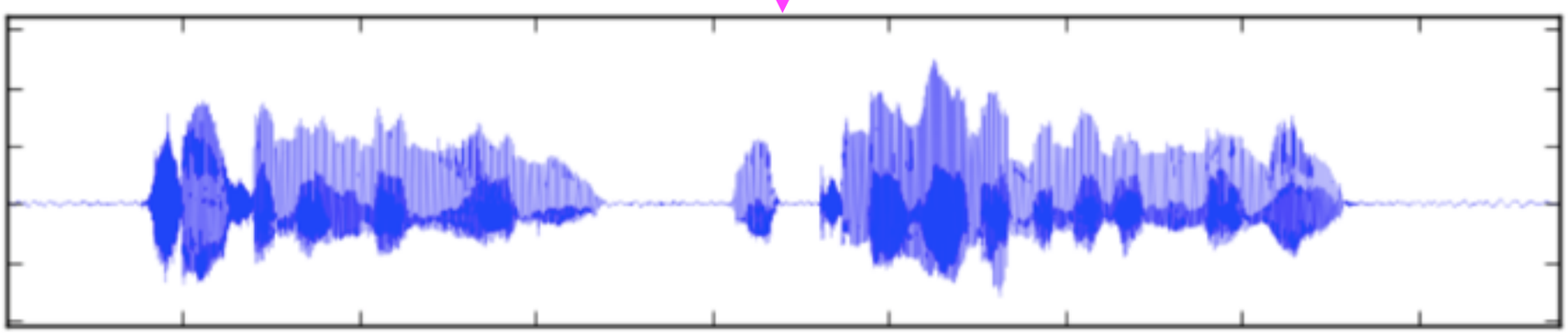
言語学・
自然言語処理

言語情報と音響的特徴量の対応付け

機械学習

音声の音響的特徴量抽出

信号処理



言語的情報を分析する適切な単位とは

- パッと思いつくのは単語
- 辞書に含まれている全ての単語を異なる単位と考えるとどうなるか？
    - すべての単語を読み上げた音声収録が必要
    - 世界最大の英語辞書オックスフォード辞書の場合
        - 29万単語
        - 1単語＝１秒だとしたら、29万秒
        - 81時間 (休憩や単語と単語の間の間を含めた場合１０倍以上の収録時間)
        - 81枚分のCD
        - 56.7ギガバイト (CD１枚=700メガバイト)
- 不可能ではないが、非効率
- 単語よりも小さい単位が適切：Sub-word単位

音素

-声調言語（例：中国語）を除いては、通常sub-word単位として「音素」を使用

-音素：２つの異なる単語の意味を区別することが可能な「音声」の最小単位

- 例：「滝」と「柿」（/a/ /k/ /i/は共通、先頭の子音/t/と/k/が異なる）
- 子音/t/と/k/のみで単語の意味を区別可能→ /t/と/k/は音素である

-異なる単語でも共通音素がある

-つまり、音素単位なら全ての単語を読み上げて音声収録する必要がない

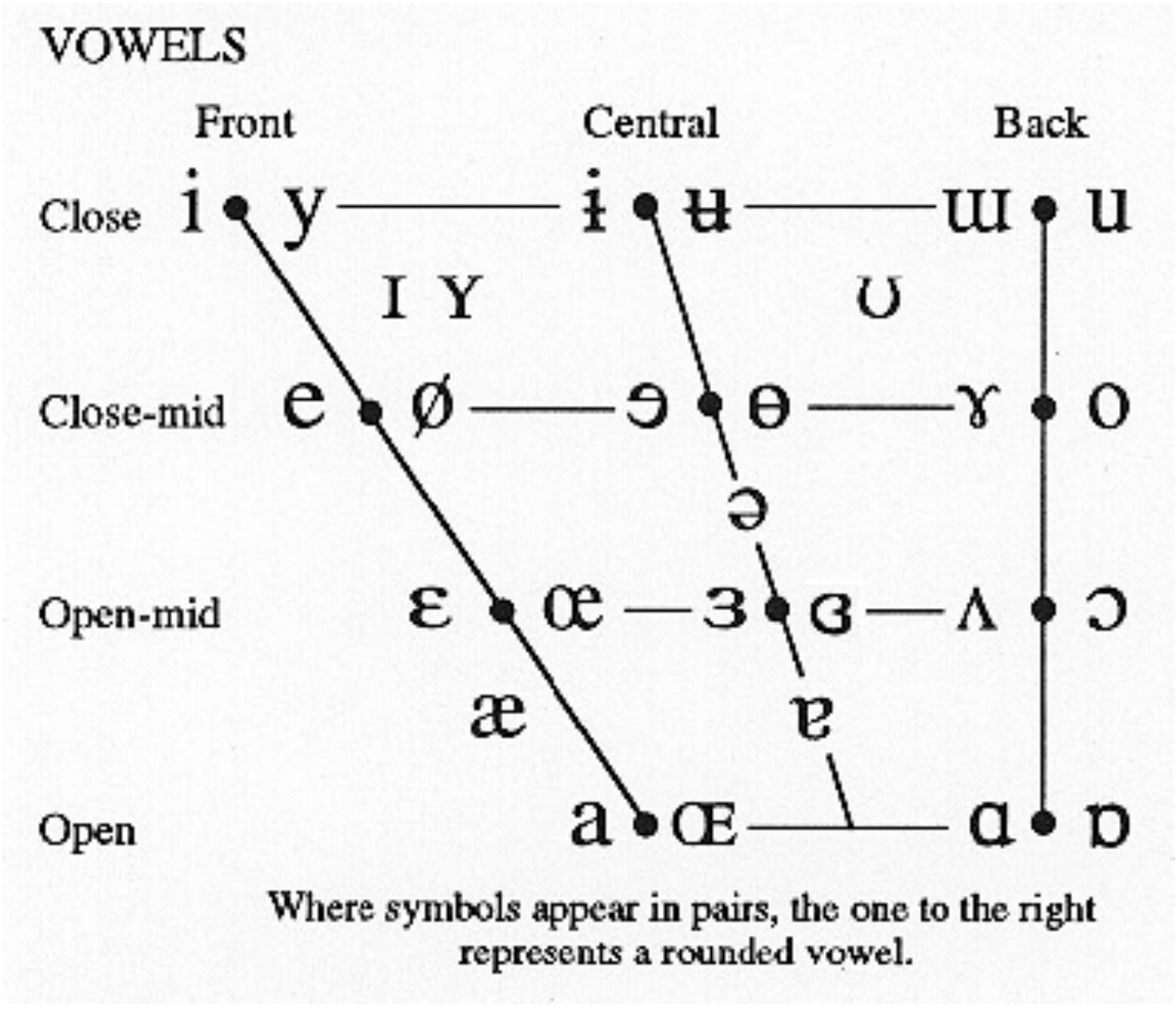
大

小

	日本語	例
文		赤い空
句	アクセント句	あかいそ｜ら
語	いわゆる単語	赤い 空
音節	個々のひらがな	/a/ /ka/ /i/ /so/ /ra/
音素	個々の母音・子音	/a/ /o/ /i/ /k/ /s/ /r/

世界の母音

国際音声記号もしくは**International Phonetic Alphabet** (IPA)



https://ja.wikipedia.org/wiki/国際音声記号より引用

世界の子音のIPA記号

THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993)

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q G		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	N		
Trill	B			r					R		
Tap or Flap				ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

https://ja.wikipedia.org/wiki/国際音声記号より引用

音素セット

- 日本語の音素セットの例 (Wikipediaから引用)

母音	/a/, /i/, /u/, /e/, /o/
子音	/k/, /s/, /t/, /c/, /n/, /h/, /m/, /r/, /g/, /z/, /d/, /b/, /p/
半母音	/j/, /w/
特殊モーラ	/N/, /Q/, /H/

- 英語の音素セットの例（音声合成ソフトFestivalでの例）
    - 15母音（二重母音含む）
    - 23子音　(6 stops, 10 fricatives, 3 nasals, 4 liquids)
    - 約40音素
    - 60音素と定義する方言もあり

-これらの音素セットは方言によって異なる

発音辞書「レキシコン」

-日本語では単語から音素への変換は容易

-英語では単語から音素への変換は、決定論的に行えない

　　-７割程度の単語しか文字から音素を推論不可

　　-単語と音素の対応付けを辞書化：レキシコン

-英単語のレキシコンの例

a	ax
abbreviate	ax b r iy v iy ey t
ability	ax b ih l ix t iy
able	ey b el
ably	ey b l iy
about	ax b aw t
above	ax b ah v
abruptly	ax b r ah p t l iy

その他の情報も付与することもある

　　英語の強勢（ストレスマーク）

その他の言語情報

- 形態素解析
  - 名詞や動詞といった品詞
  - 単語のかかり受けや句の構造解析
- 形態素解析の結果をもとに更に多くの言語的情報を抽出
  - アクセント句境界・アクセント核位置推定
  - ポーズや呼気の挿入位置決定

2006 年の調査によると，日本全国で，約 33%の家庭がペットを買っているそうです。
ニセ’ ン/ロク’ ネンノ/チョ’ ーサニヨルト_ニホンゼ’ ンコクデ_ヤ’ ク%/サ’ ンジュー/
サンパーセ’ ントノ/カテーガ/ペ’ットオカッ.テイルソ’ ーデス%.
グッズの専門店もでき，お洒落な服を着た犬も，よく見かけます。
グ’ッズノ/センモ’ ンテンモ/デ’ キ_オシャ’ レナ/フ%ク’ オ/キ%タイヌモ_ヨ’ ク/ミカケマ’ ス%.

’:アクセント核，/:アクセント句境界，_:ポーズ，%:母音の無声化

- JEITAフォーマットによる表記例（峯松らの論文から引用）

音声合成：文章と音声の対応付け（マッピング）

小さな鰻屋に，熱気のようなものがみなぎる

文章の言語的情報抽出

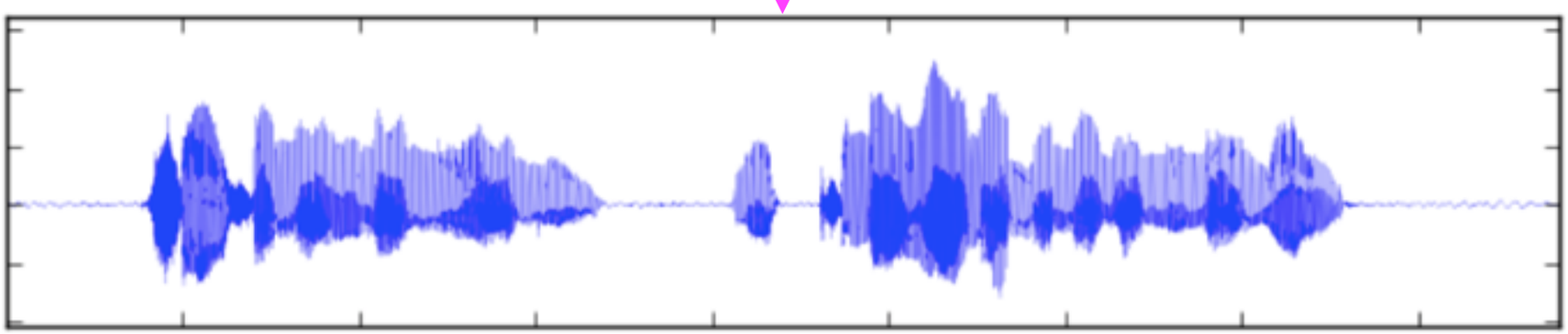
言語学・
自然言語処理

言語情報と音響的特徴量の対応付け

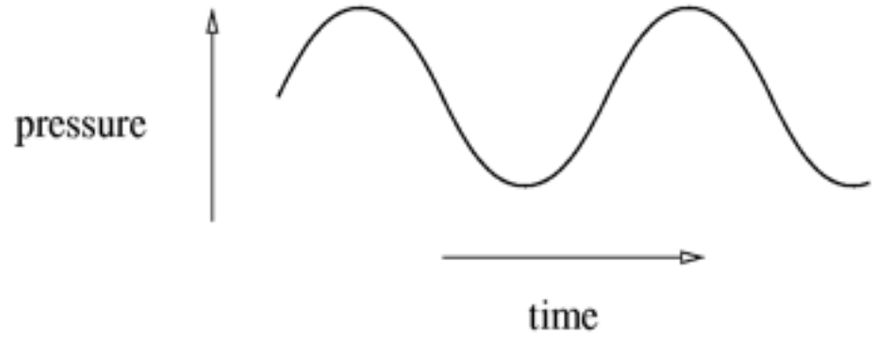
機械学習

音声の音響的特徴量抽出

信号処理



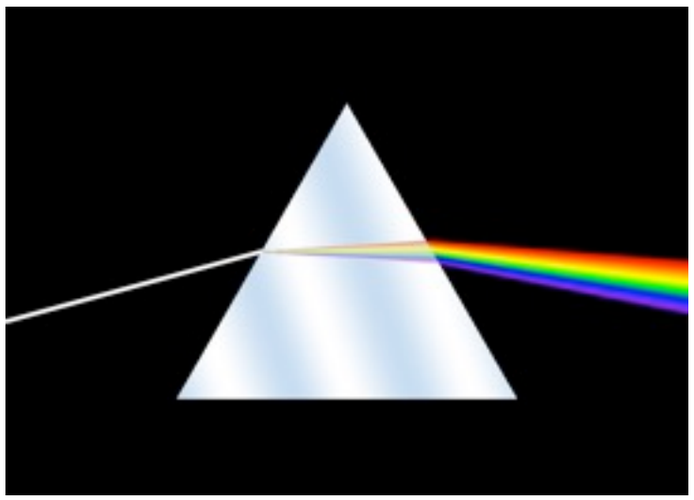
音声波形の周波数変換：スペクトル表現

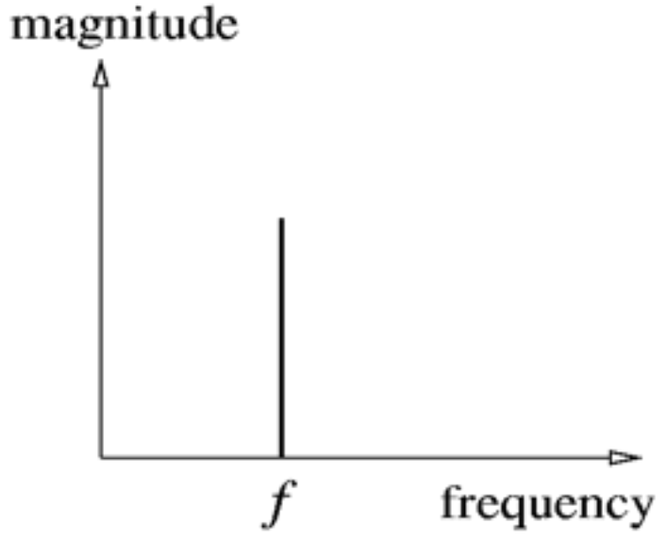


音声波形は１次元の信号

フーリエ変換等により波形信号を周波数表現したもの：スペクトル

フーリエ変換のたとえ：光（波形）を色（周波数）に変換するプリズム

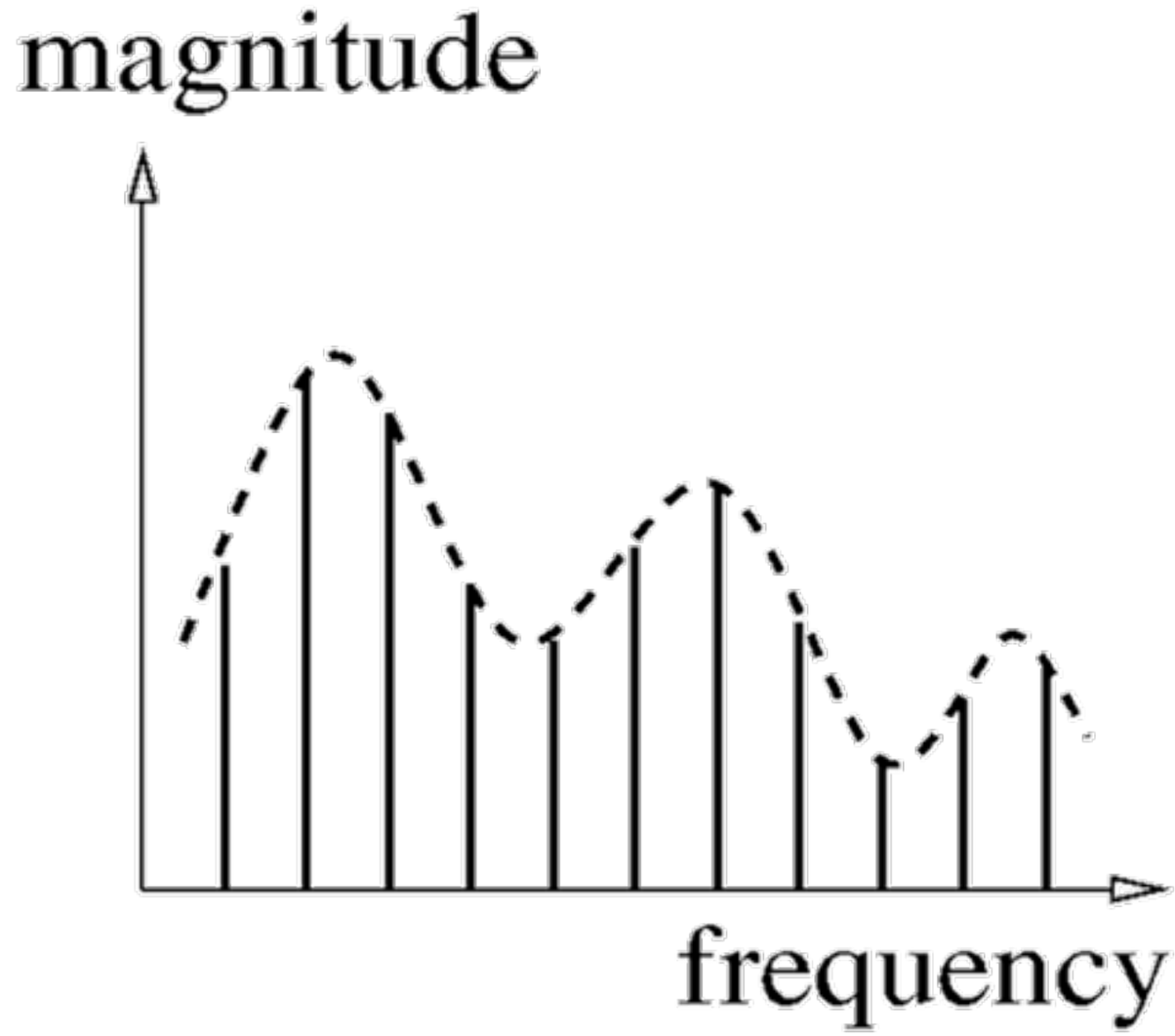




サイン波の場合、特定の周波数のみピークが出現

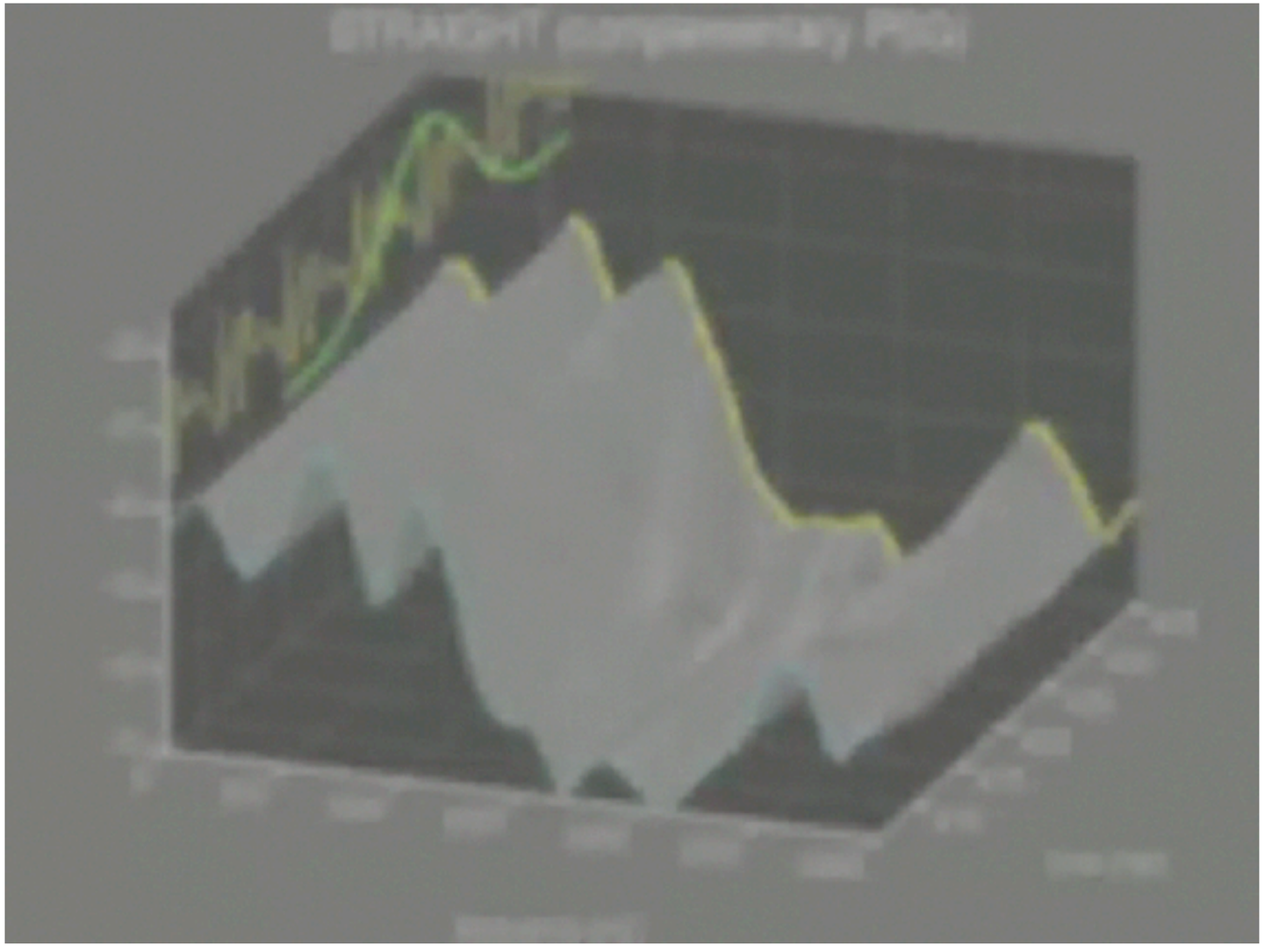
音声のスペクトル表現はどのように表現されるか？

有声音のスペクトル



母音のスペクトル

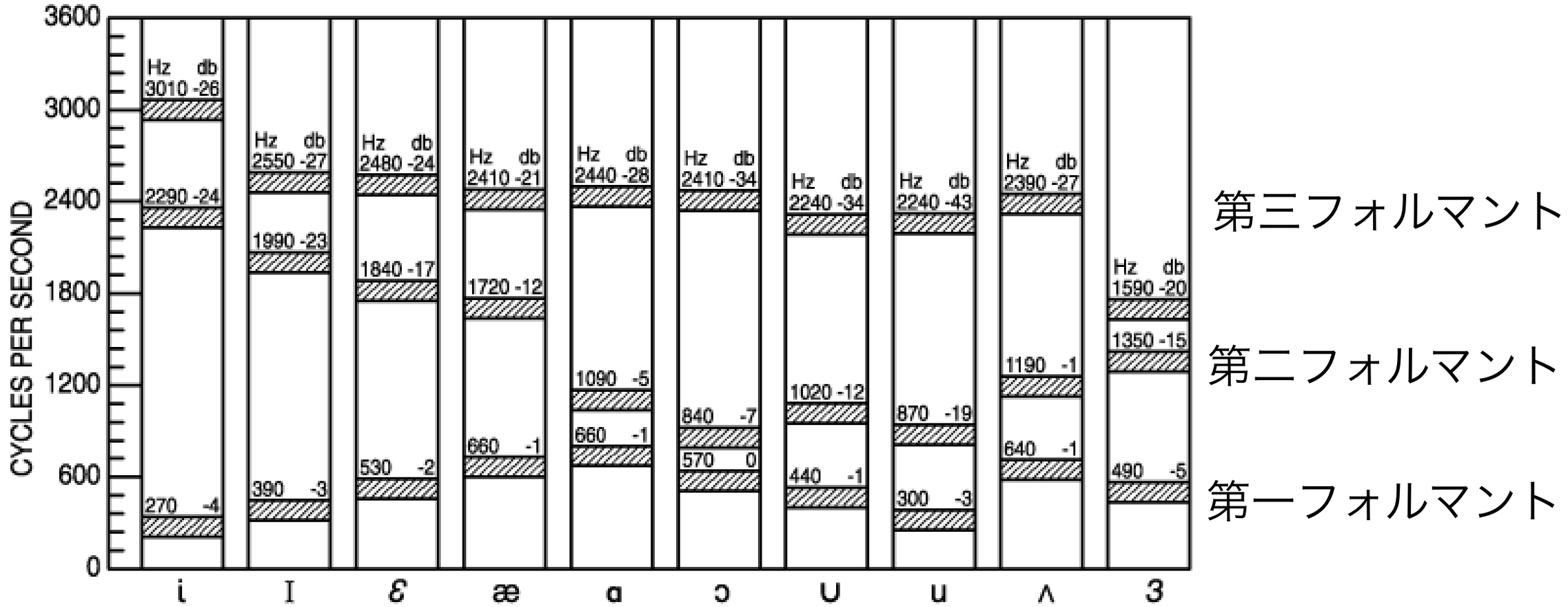
- 母音は色々な周波数成分が混合
- ２～３の周波数成分が強いピークを示す
- **フォルマント周波数**（声道の共振周波数）：低域から第一、第二、第三と呼ぶ



- 和歌山大学河原教授により作成

各母音のフォルマント周波数

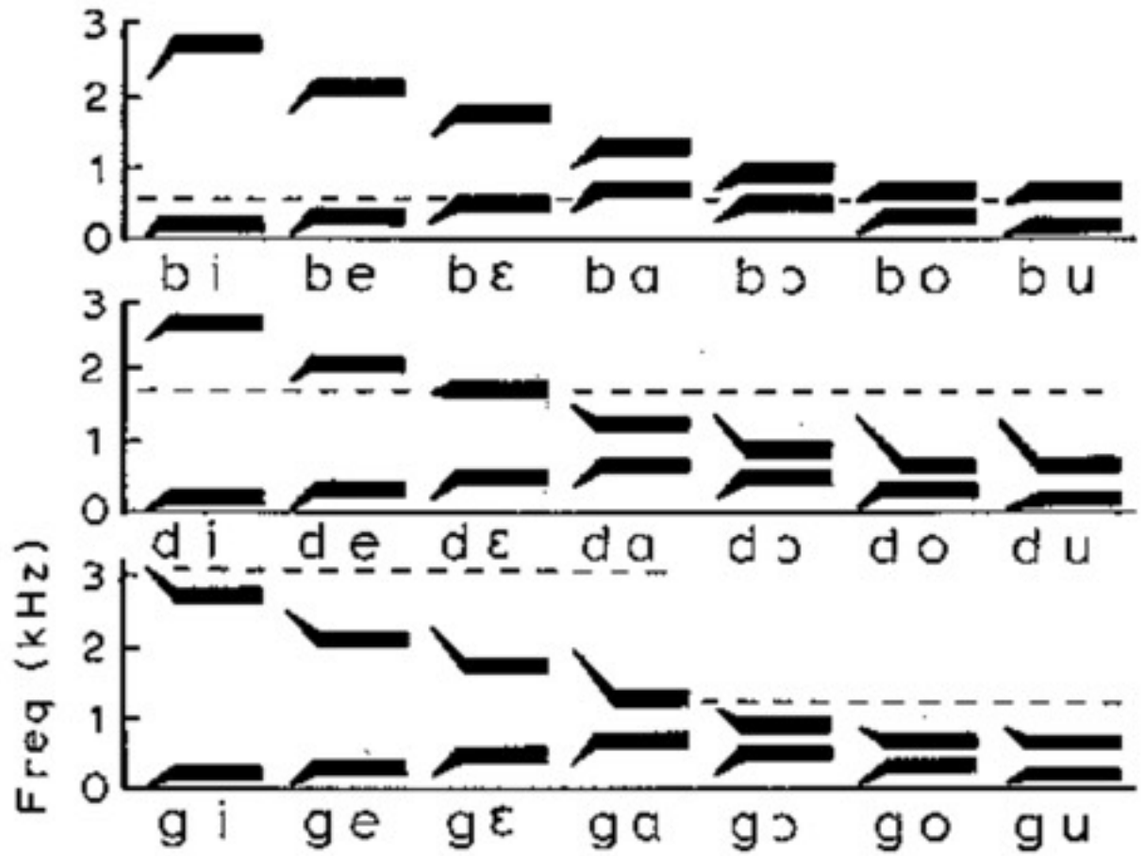
- **母音は第一フォルマント周波数、第二フォルマント周波数**（およびしばし第3フォルマント）により特徴付けられる



引用：Jont B. Allen James L.Flanagan and Mark A.Hasegawa-Johnson, Speech Analysis Synthesis and Perception, Springer-Verlag, 2008.

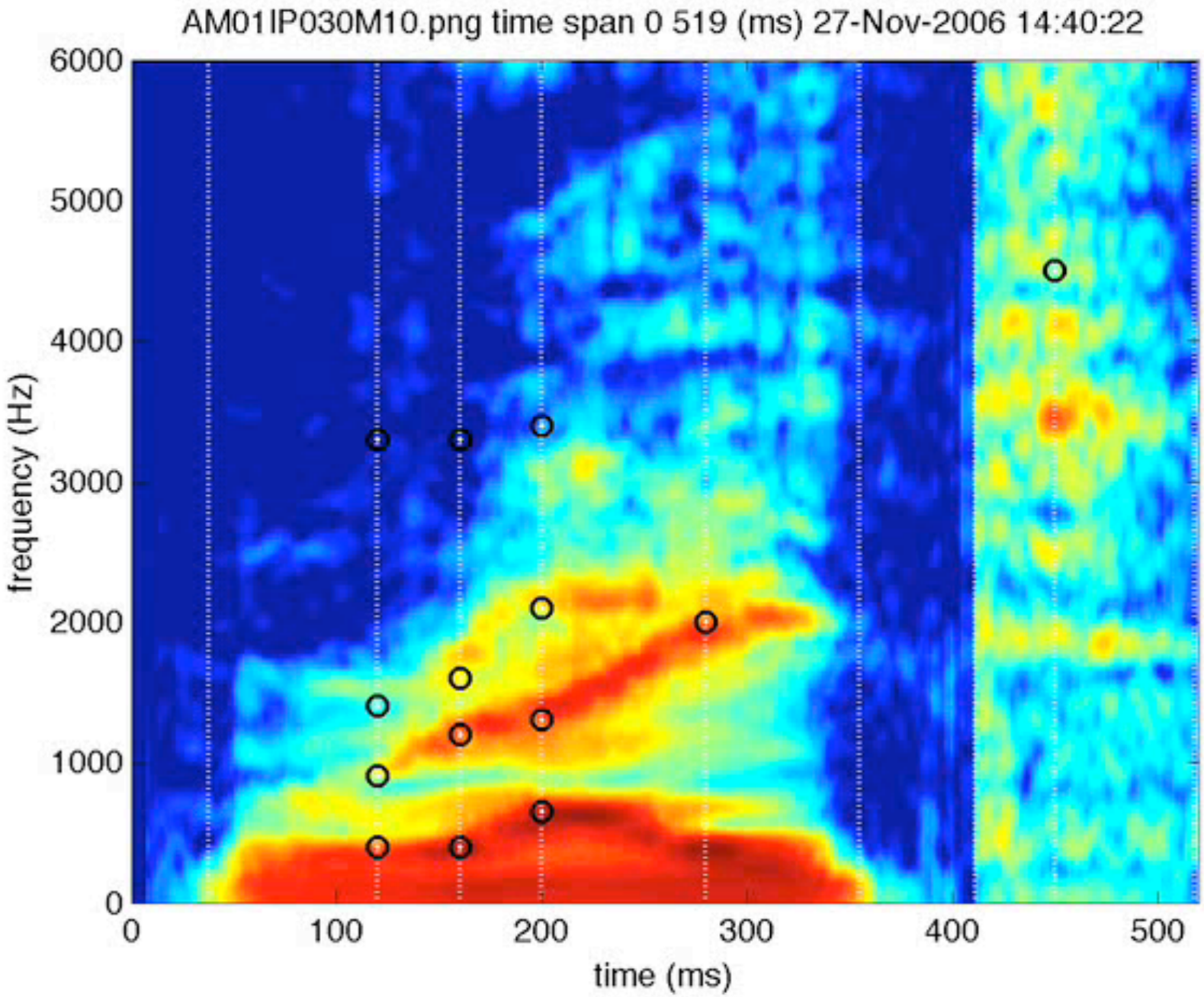
子音の場合： /b/, /d/, /g/ の場合

- 幾つかの子音は、“**フォルマント遷移**”により特徴付けられる
- /b/ /d/ /g/ “**第二フォルマント遷移**”



Delattre, P. C., A. M. Liberman and F. S. Cooper (1955) Acoustic Loci and Transitional Cues for Consonants. JASA vol. 27, no. 4. 769-773.

子音 /l/, /r/ の場合：第三フォルマント

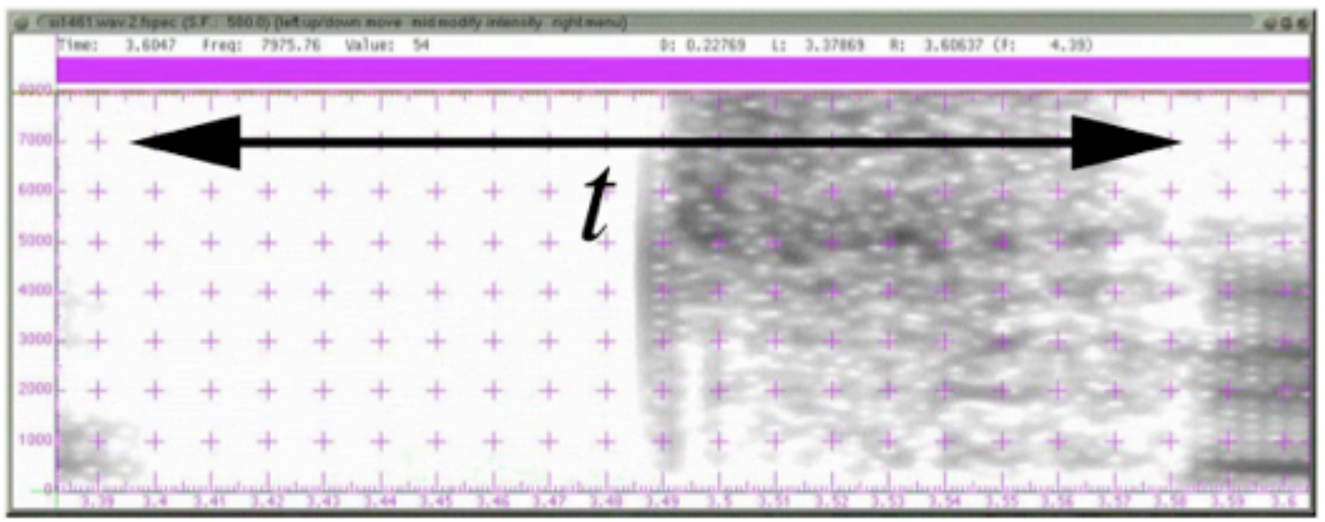


/right/ -> /light/

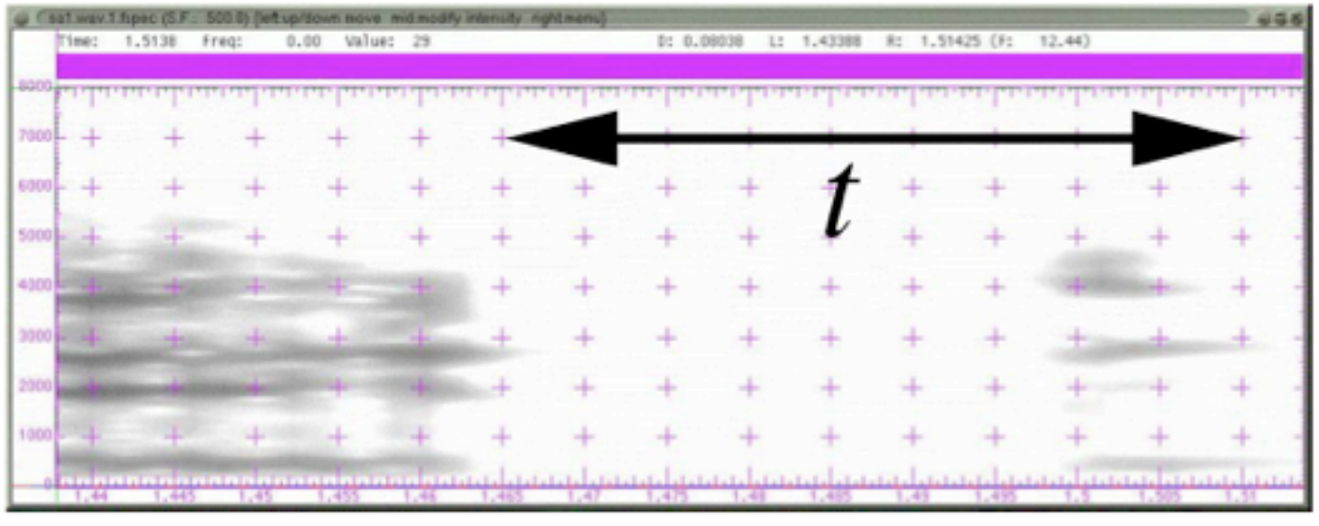
- 和歌山大学河原教授により作成

しかし、、

フォルマントが全ての音素の特徴を説明できる訳ではない



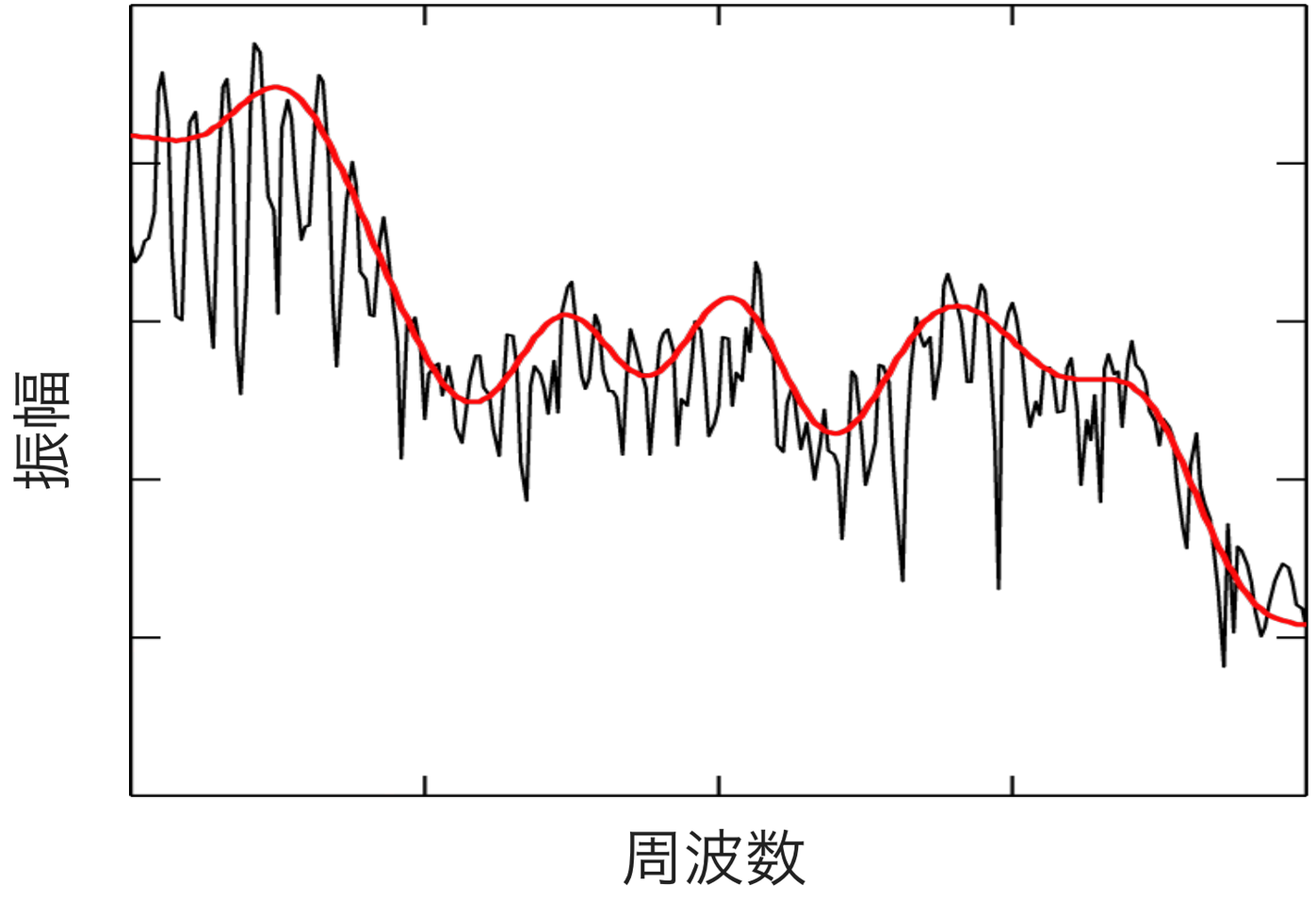
"tube"



"suit"

スペクトル包絡抽出

- フォルマントは、音素との対応関係が非常に明確
- しかしフォルマントは音声のスペクトルの微細構造を再構築する上で不十分
- 通常、フォルマント周波数の情報だけでなく、周波数全域のスペクトルの概形（**スペクトル包絡**）を抽出する線形予測分析やケプストラム分析を利用



ケプストラム分析により得られらたスペクトル包絡

音声合成：文章と音声の対応付け（マッピング）

小さな鰻屋に，熱気のようなものがみなぎる

文章の言語的情報抽出

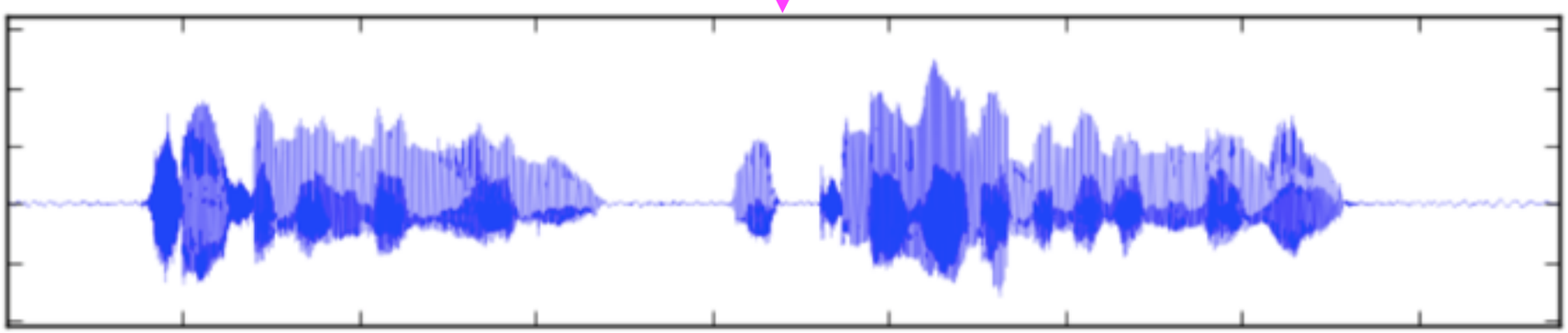
言語学・
自然言語処理

言語情報と音響的特徴量の対応付け

機械学習

音声の音響的特徴量抽出

信号処理



言語情報と音響特徴量の対応付け

- 音素等の言語情報と音響特徴量もしくは音声波形とをどう対応づけるかにより、色々な音声合成手法が開発された
  - フォルマント合成：
    - 音素とフォルマントとの対応関係を決定論的に実装
  - ダイフォン合成：
    - 音素（ダイフォン）と音声波形を対応付け
  - 波形接続合成：
    - コンテキスト依存音素と音声波形を対応付け
  - 統計的音声合成：
    - コンテキスト依存音素とスペクトル包絡との対応付けを確率的に学習

フォルマント合成

- フォルマント合成システムの生成手順
  - 文章を音素へ変換
  - 音素情報をもとにフォルマント周波数 (およびその他の変数)を決定
  - 決定されたフォルマント周波数をもとに音声波形を合成

- 代表的なシステム
  - Parametric Artificial Talker (PAT) (Walter Lawrence, Edinburgh, 1950s-1960s)
  - OVE (Gunner Fant, Sweden 1960s)
  - MITalk (1970s), KLATTalk, DECTalk

Example 1: MITalk

ダイフォン合成

- ダイフォン合成の生成手順
  - 個々のダイフォンの音声波形を予め**１つずつ**用意
    - ダイフォン：音素の後半部と後続音素の前半部から構成される音素単位
  - 文章をダイフォン系列へ変換
  - 個々のダイフォンに該当する音声波形を接続
  - 信号処理を適用し、平滑化および韻律処理

- 代表的なシステム
  - Sparte (Courbon and Emerald, France telecom, ’82)
  - ボーカロイド（Yamaha, Japan, 2003）

Examples: diphone synthesizer

Sparte: A text-to-speech machine using synthesis by diphones
Courbon, J.-L.;　Emerard, F.; Proc. ICASSP-82

波形接続合成

- 波形接続合成の生成手順
  - 個々の音素の音声波形を**種々のコンテキストを考慮し複数**用意
    - コンテキスト：前後の音素など音響的に影響を与える要因
  - 文章をコンテキスト依存音素系列へ変換
  - 目標のコンテキスト依存音素系列に完全一致もしくは部分一致する複数の候補の中から最適な音声波形を動的計画法により選択・接続

- 代表的なシステム
  - CHATR (Hunt and Black, ATR, Japan, ’95)
  - Festival (Black, CSTR, Edinburgh, UK, ’97)
  - AT&T Natural voice (USA)

Examples: unit selection synthesizer

統計的音声合成

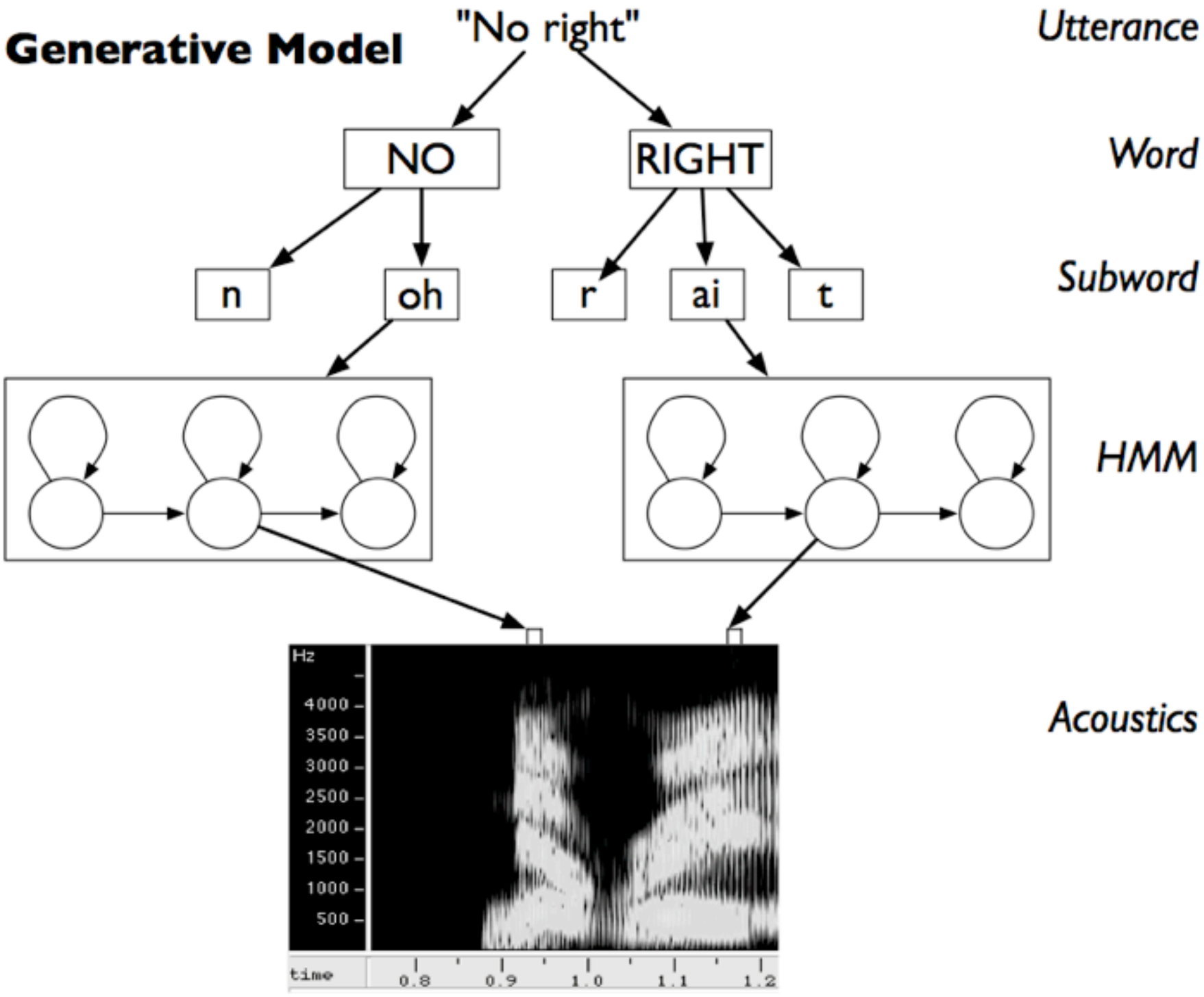
- 統計的音声合成の手順
  - 学習フェーズ
  - 個々の音素の音声波形を種々のコンテキストを考慮し複数用意
  - **個々のコンテキスト依存音素の特徴を隠れマルコフモデル（Hidden Markov Model, HMM）などの時系列統計モデルにより学習**

- 合成フェーズ
- 文章をコンテキスト依存音素系列へ変換
- **時系列統計モデルから最適音響特徴量を予測**
- 予測音響特徴量から音声波形を合成

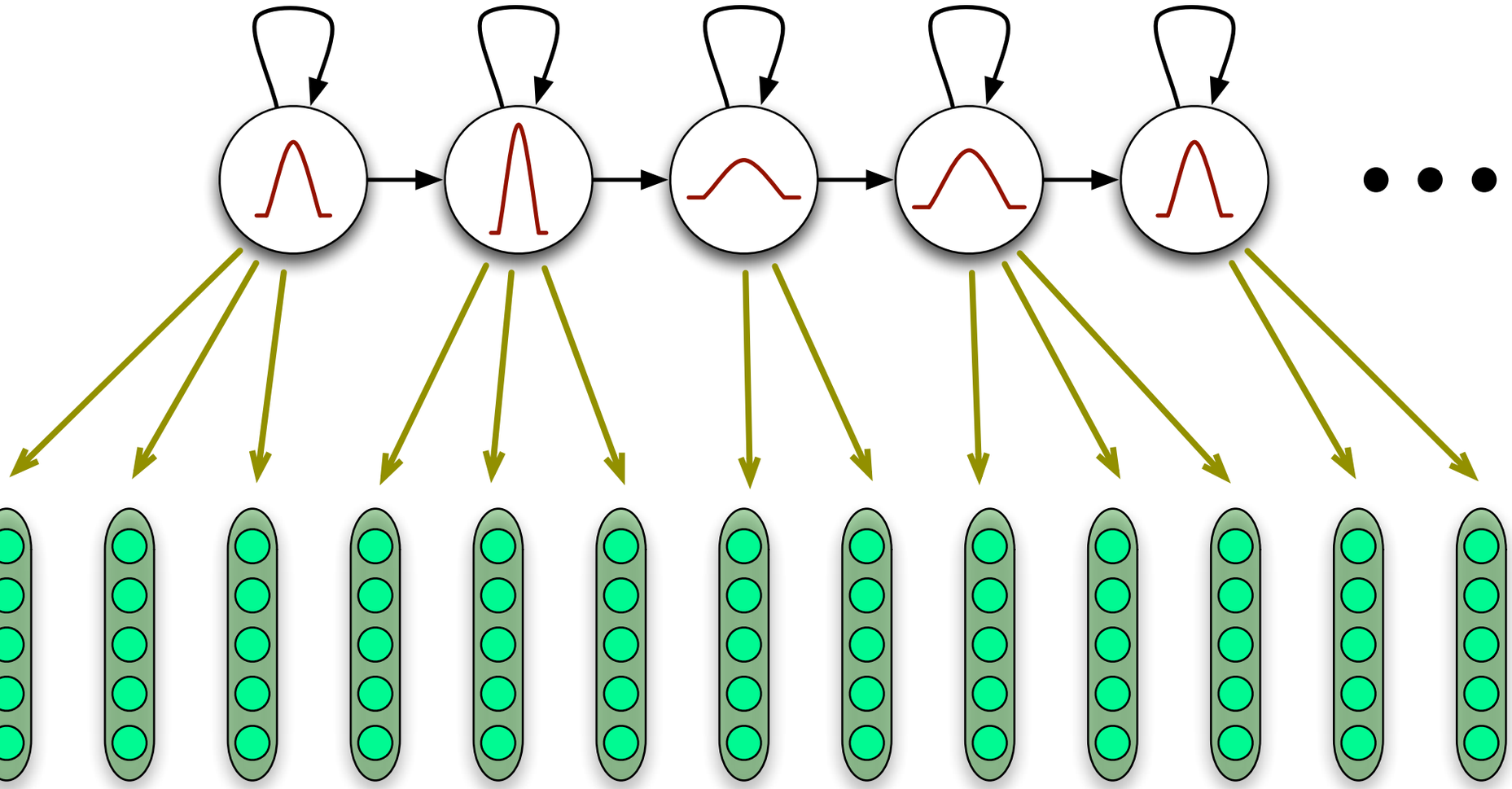
Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Keiichi Tokuda,
“The HMM-based speech synthesis system (HTS) version 2.0,” SSW 6, pp.294-299, Aug. 2007.

Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura “Speech Synthesis Based on Hidden Markov Models” Proceedings of The IEEE, 2013

言語情報と音響的特徴量の確率的対応付け



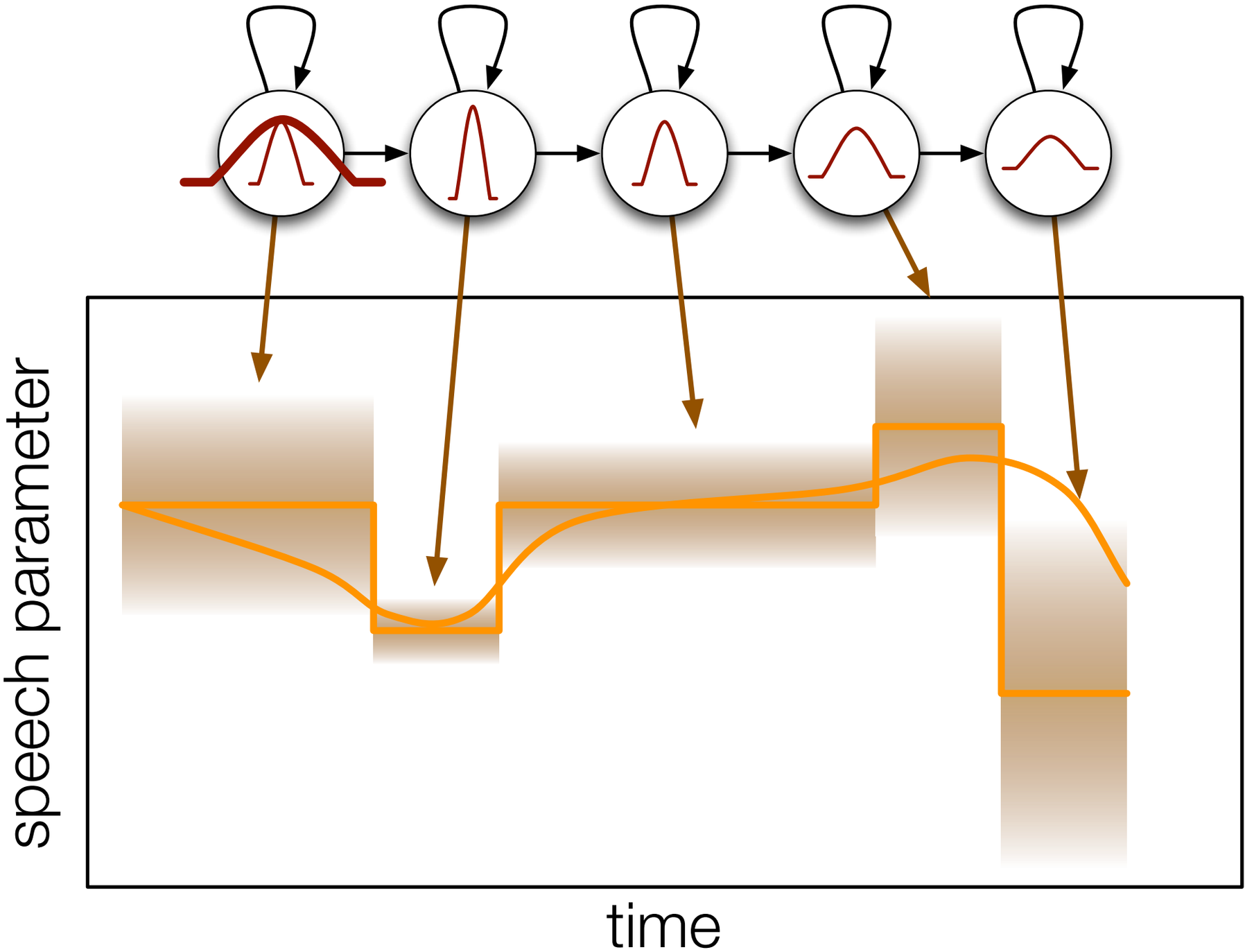
HMMの概念的説明



例：名詞の中の、前の音が子音「r」、次の子音が「t」である母音「ai」の音響モデル

- 各単位を幾つかのセグメントに分ける
- 境界を勝手に決めてくれる
- そのセグメント内の変動を分布で表す

統計モデルからの音声パラメータの生成



HMMに基づく音声合成

- 名古屋工業大学が中心となってオープンソースで無償公開
  - HTS (H Triple S 2002年以降〜)
  - 音声情報処理に関する有名な国際学会Interspeech 2012
    - 76%の音声合成論文がHTSを使用
- 産業応用も進む（一例）
  - 株式会社NTTドコモの携帯電話12機種への搭載
  - HOYAサービス株式会社の「VoiceText Micro SDK」への搭載
  - Nuance Communication社（米国）の「Nuance Vocalizer」への搭載
  - 株式会社KDDI研究所の「N2 TTS」及び「ささやくヤーツ」への搭載
  - Google社（米国）の「Android OS」への搭載

統計的音声合成の応用

- 統計的音声合成は様々な制御が可能であり幾つもの新規アプリを生み出した
  - 話者適応：モノマネ、パーソナライゼーション
  - モデル補間：表現の強弱
  - クロスリンガル音声合成
  - 騒音下でも聞きやすい音声合成
  - 障害・福祉応用

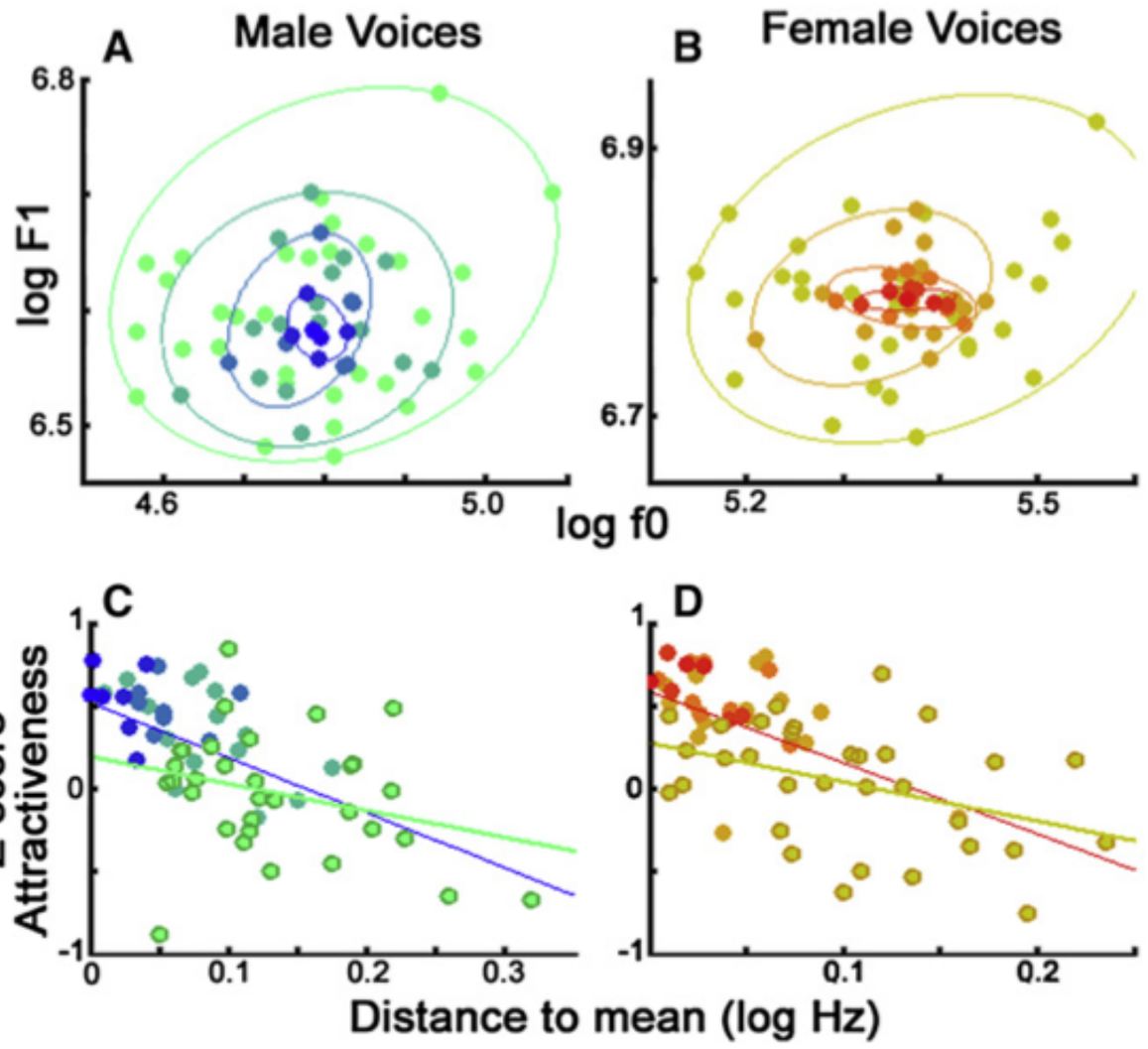
統計モデルを利用するとその他何が嬉しいのか

- 音声をすべて関数で表現可能
  - 音声波形は、関数表現を自動で学習する際に利用
  - 音声の合成には、推定された関数を利用
  - たった数MBという音響モデル（関数の集合）で声を合成できる
    - 携帯やタブレットといった端末でも高品質な音声合成
- 学習結果の声の関数を操作可能
  - 関数を操作→声を操作
  - 合成音声をちょっと怒った声へ変化
- 実際には存在しない声も創ることも可能
  - 複数の話者を用意→個々の人の声の関数を学習→関数を平均化
  - 平均声

平均声は他の人の声より魅力的

- 平均顔：平均前の個人の顔よりも魅力的に見える
- 平均声も魅力的に聞こえます
- Pascal Belin： “Vocal attractiveness increases by averaging”, Current Biology, 20, January 26, 2010.

- 2、4、16、32名の平均声
- 基本周波数と第一フォルマント空間での平均声からの距離
- 声の魅力さのスコアとの相関係数：r= -0.59



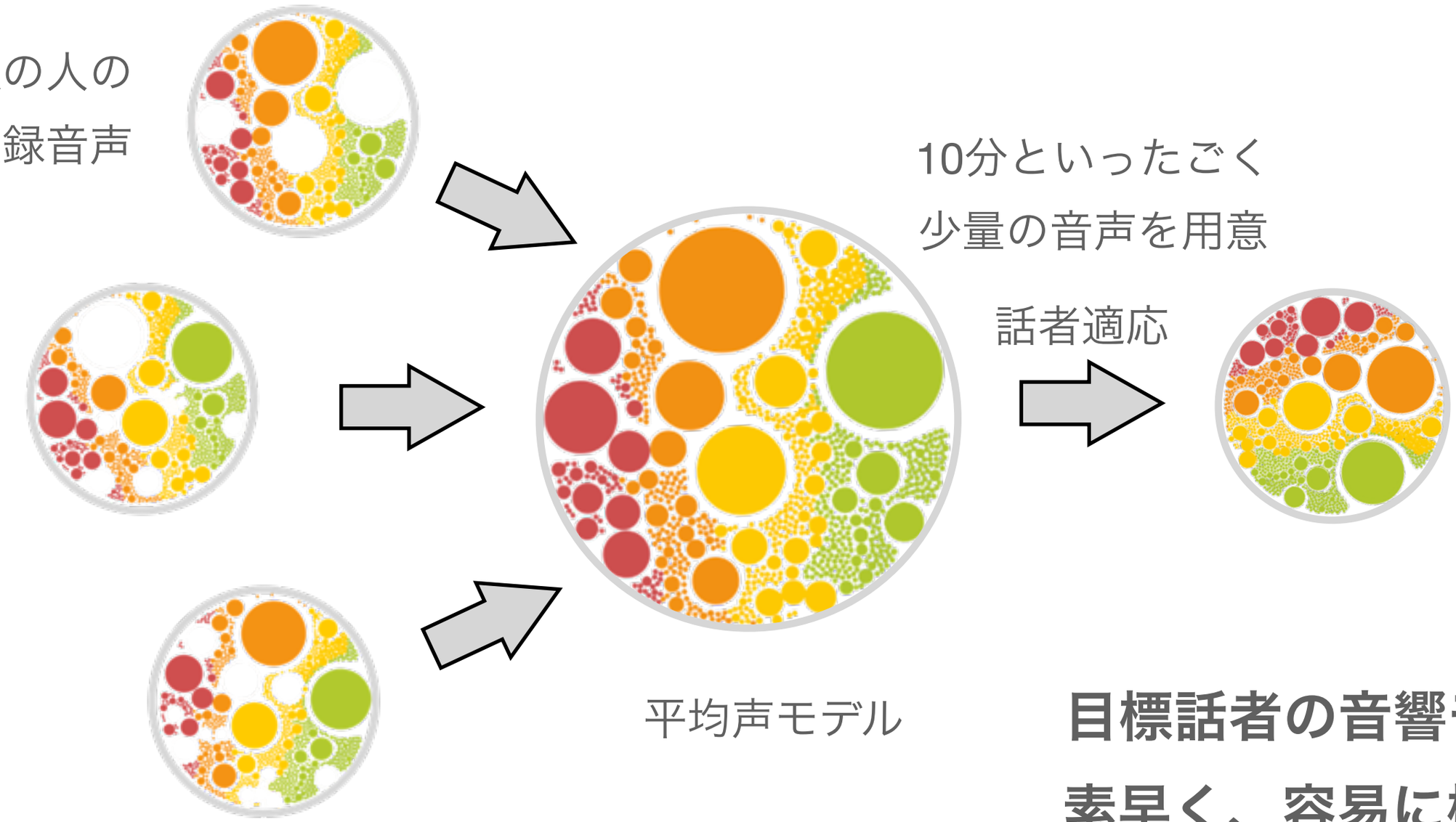
平均声を利用すると、工学的に何が嬉しいのか

複数の人の
収録音声

ある人の音響モデルをその人のデータから直接作ろうとすると、**数時間、数十時間**という音声収録が必要

平均声を利用すると、工学的に何が嬉しいのか

複数の人の
収録音声



平均声モデル

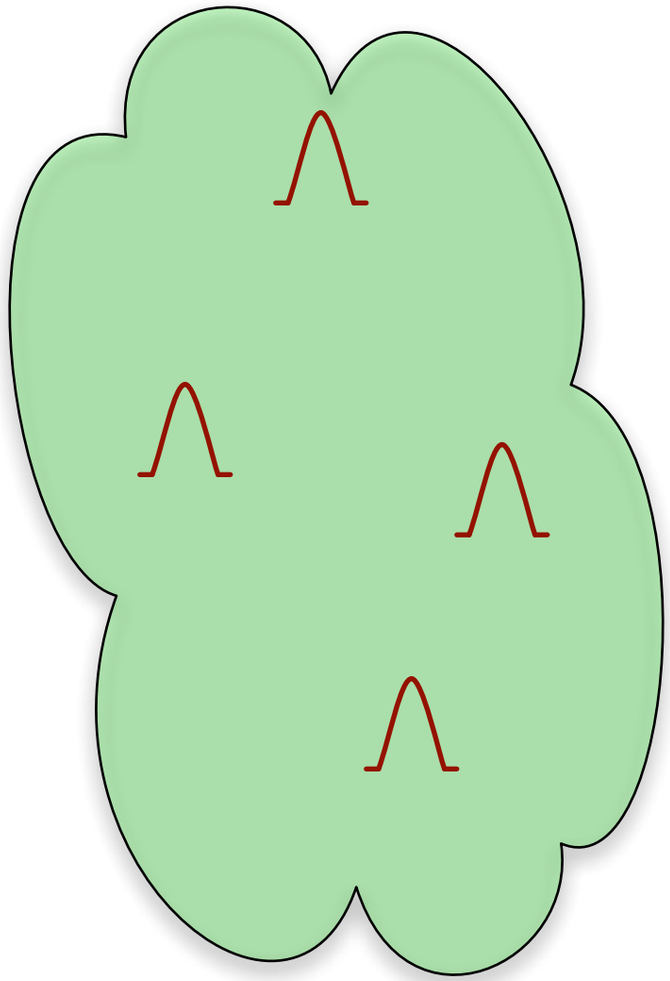
10分といったごく
少量の音声を用意

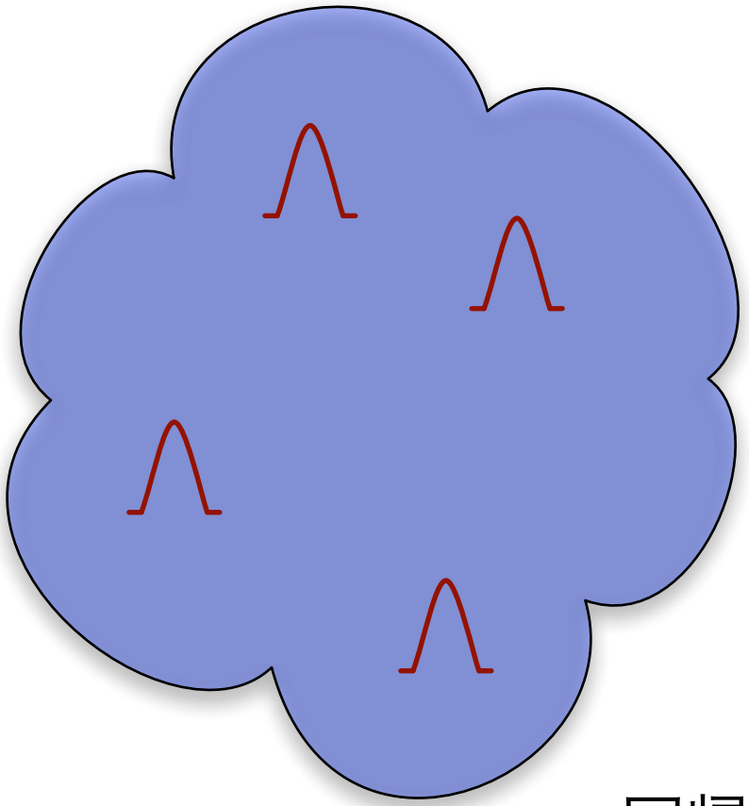
話者適応

目標話者の音響モデルを
素早く、容易に構築可能

話者適応：統計モデルのアフィン変換

回帰クラス 1



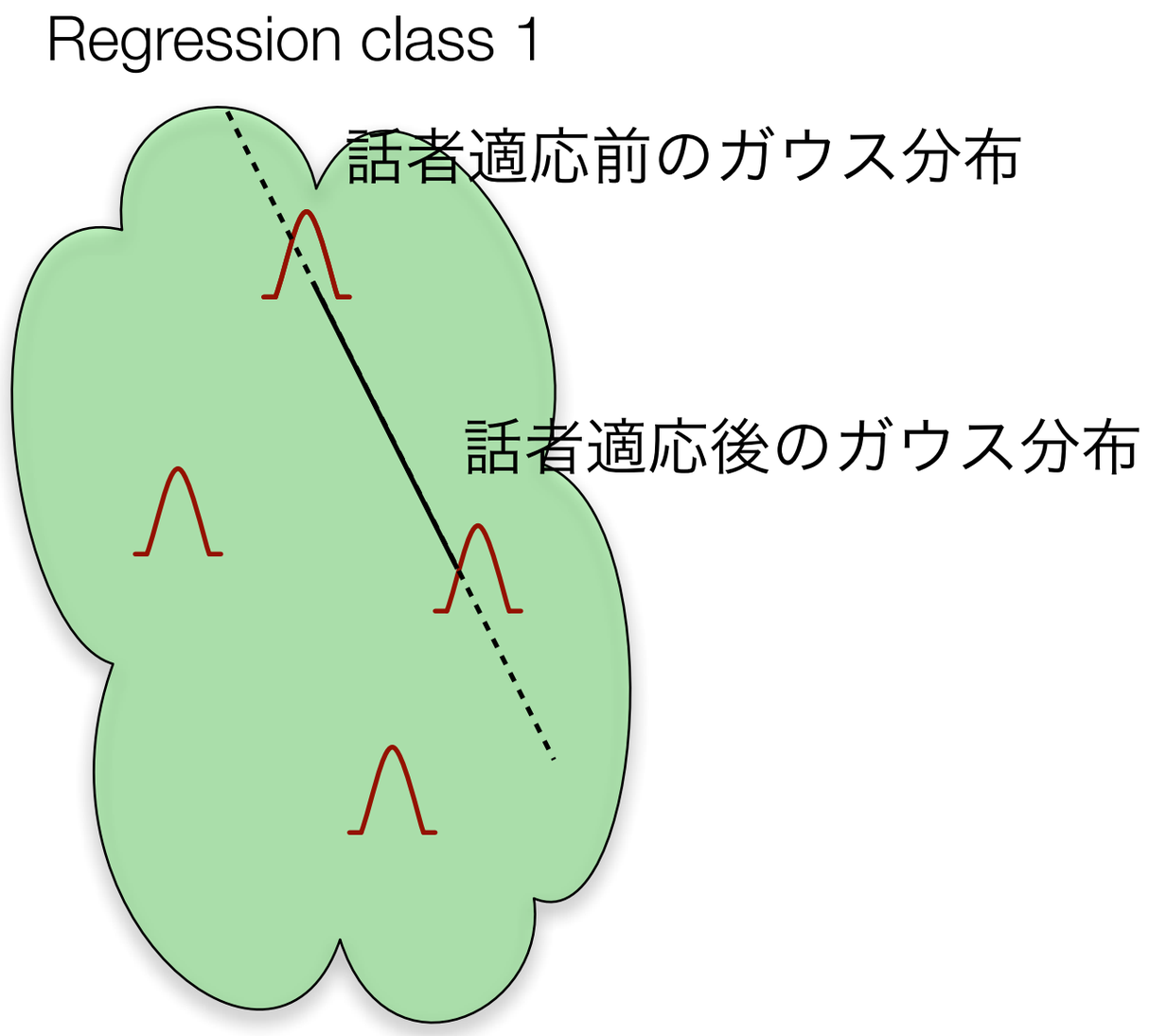


回帰クラス 2

話者適応：少量のデータで声を模倣する技術

- 問題
  - 従来の音声合成：一人あたり数十時間の音声データを収録する必要があった
  - 高コスト、限定された話者、喋り方
- HMM音声合成の話者適応技術
  - １０分ほどの少量の音声データで話者の声質を模倣することが可能
  - 最近では、良く似ているため、「声のクローン」と呼ぶ人も多い
  - どの程度似ているか示す音声サンプル
  - 低コスト

モデル補間



M. Tachibana, J. Yamagishi, T. Masuko, T. Kobayashi, “Speech Synthesis with Various Emotional Expressions and Speaking Styles by Style Interpolation and Morphing,” IEICE Trans. Information and Systems, E88-D, no.11, pp.2484-2491, November 2005.

適応および補間のサンプル

- アメリカ女性平均声から７歳の女の子への適応
  - 平均声
  - 適応結果

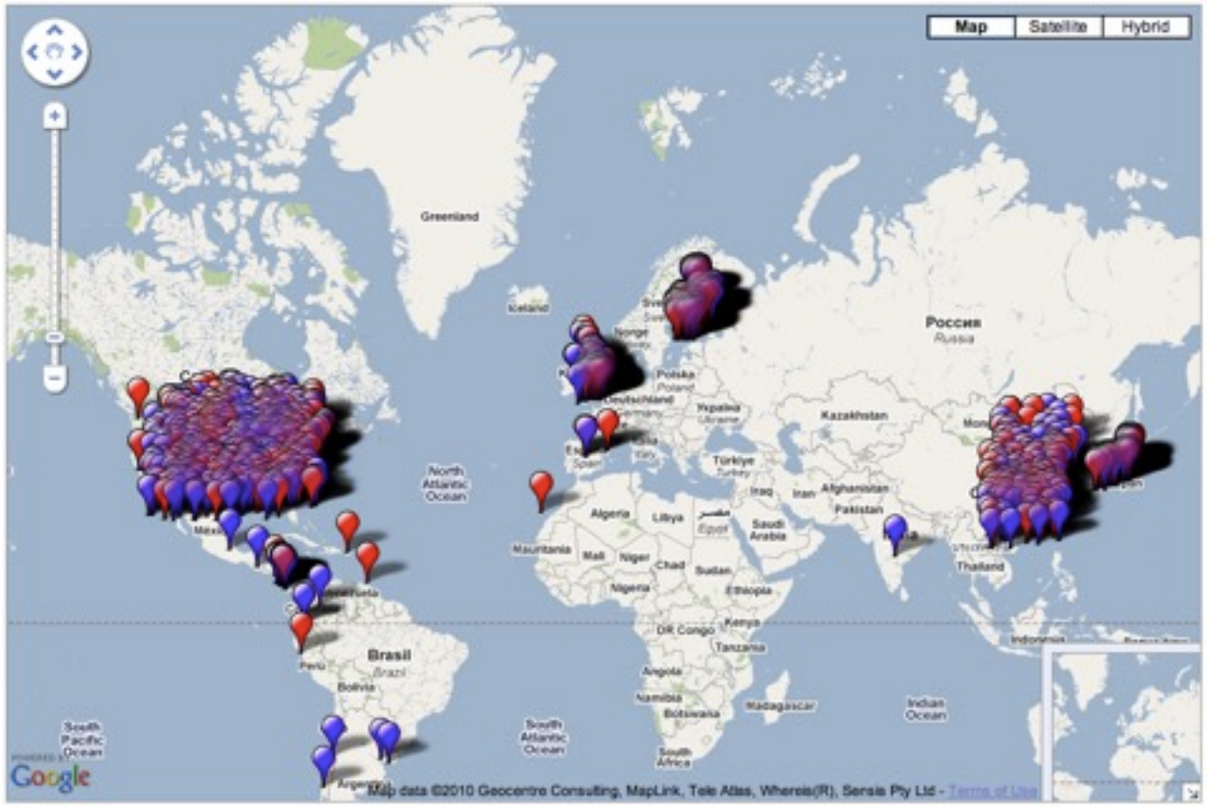
- アメリカ男性平均声からインド英語への適応
  - 平均声
  - 適応結果

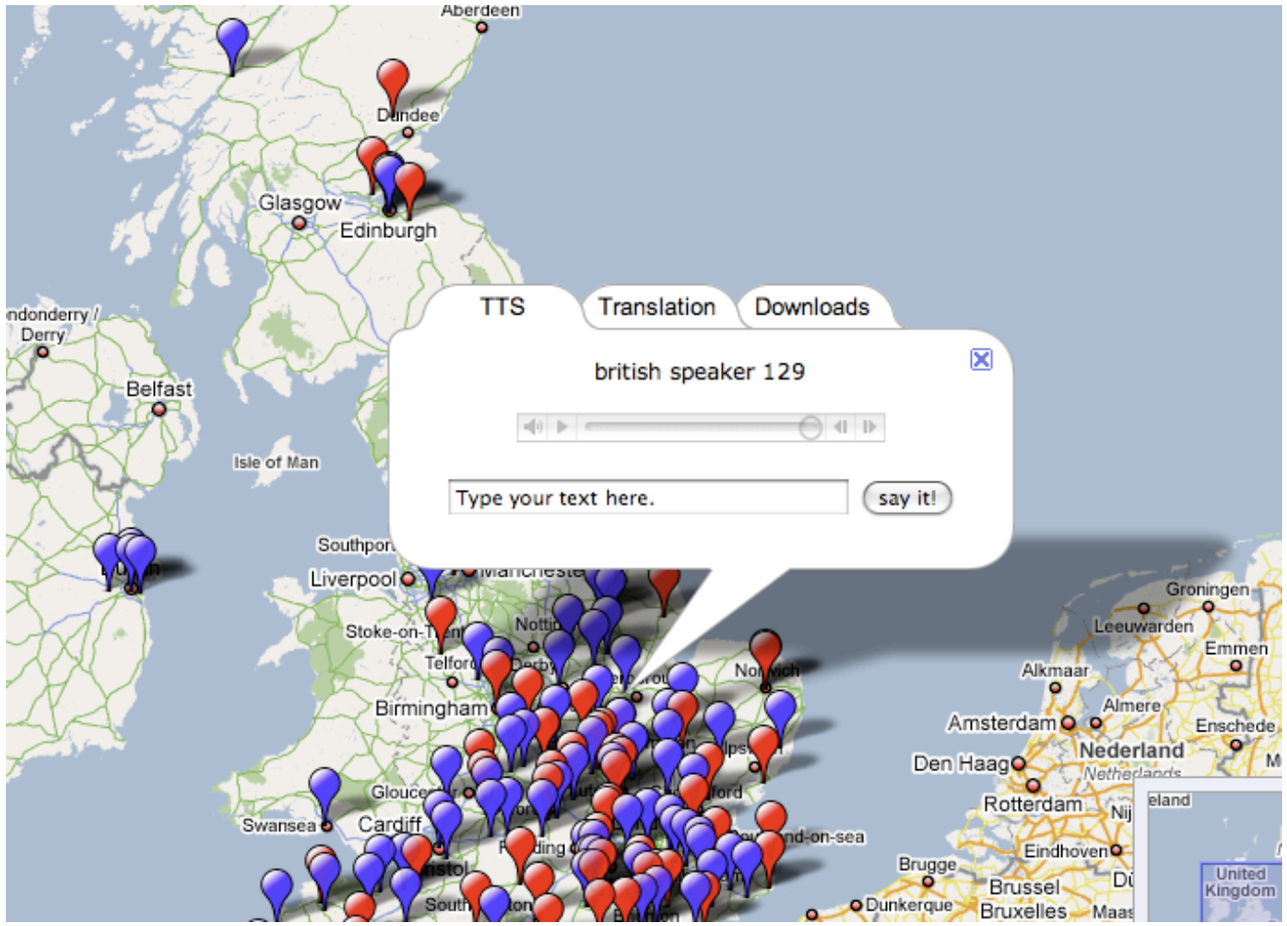
- 平静から怒りへの適応

O. Watts, J. Yamagishi, S. King, K. Berkling, “Synthesis of Child Speech with HMM Adaptation and Voice Conversion” IEEE Audio, Speech, & Language Processing, vol.18, issue.5, pp.1005-1016, July 2010

M. Pucher, D. Schabus, J. Yamagishi, F. Neubarth “Modeling and Interpolation of Austrian German and Viennese Dialect in HMM-based Speech Synthesis,” Speech Communication, Volume 52, Issue 2, Pages 164-179, February 2010

音声合成システムのパーソナライゼーション

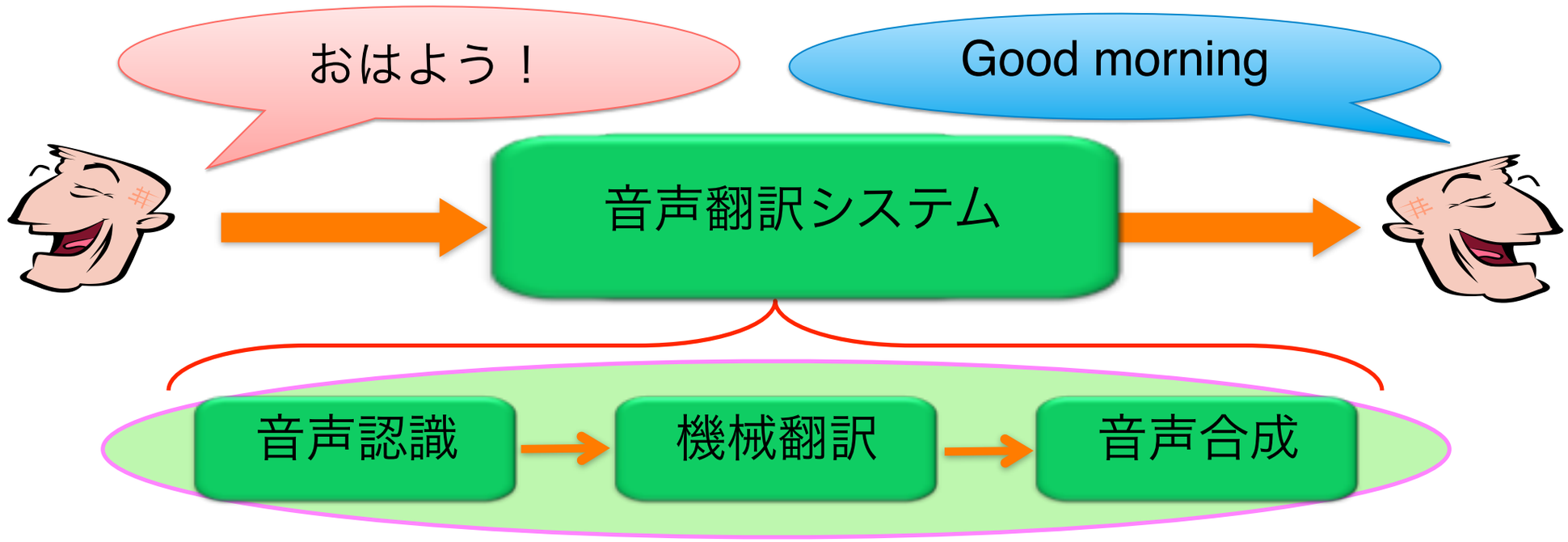




http://www.emime.org

J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, Y. Guan, K. Oura, K. Tokuda, R. Karhila, M. Kurimo, “Thousands of Voices for HMM-based Speech Synthesis -- Analysis and Application of TTS Systems Built on Various ASR Corpora,” IEEE Trans. Audio, Speech, & Language Processing, vol.18, issue.5, pp.984-1004, July 2010

自分の声で音声翻訳！

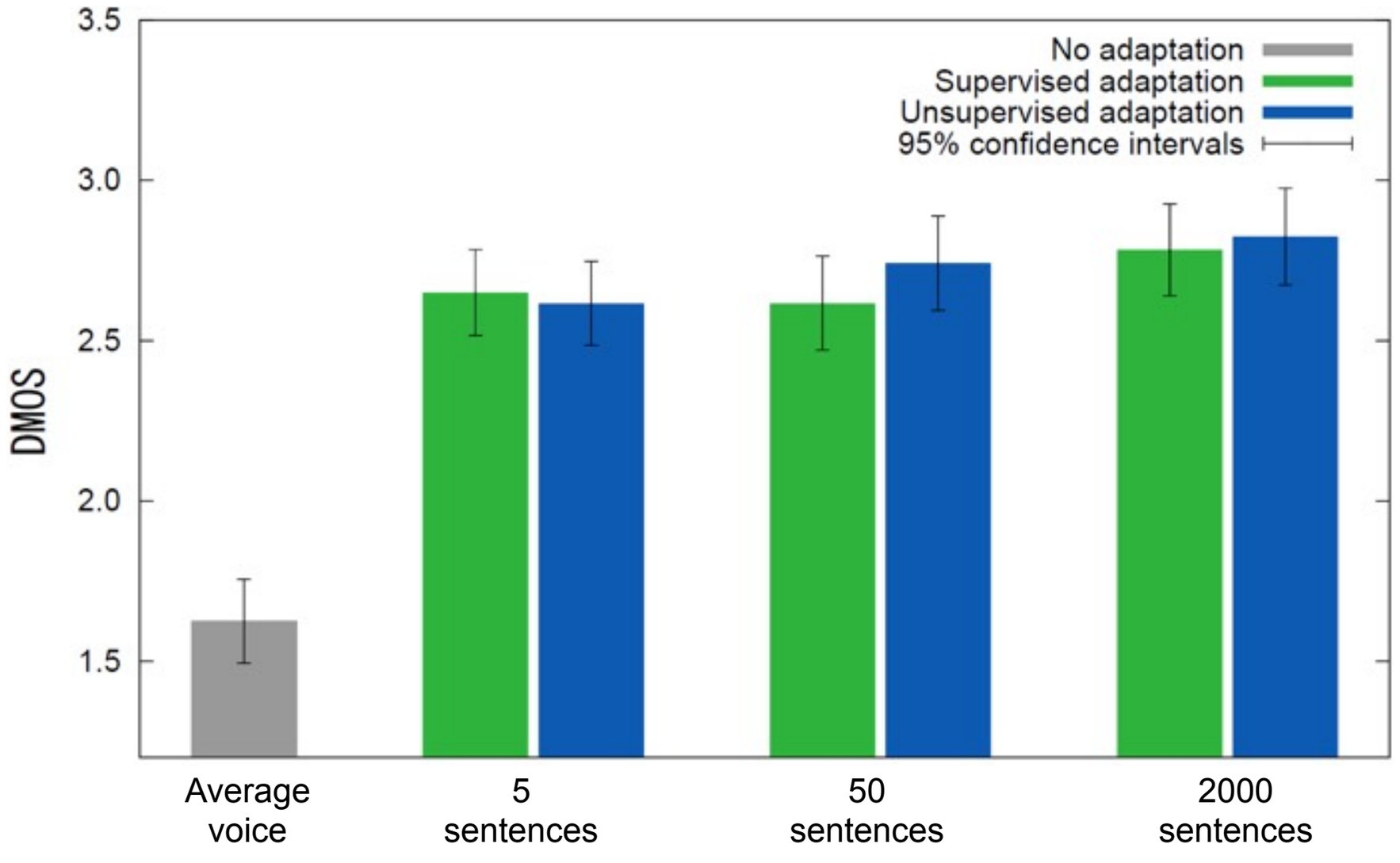


EMIME
Effective Multilingual Interaction in Mobile Environments

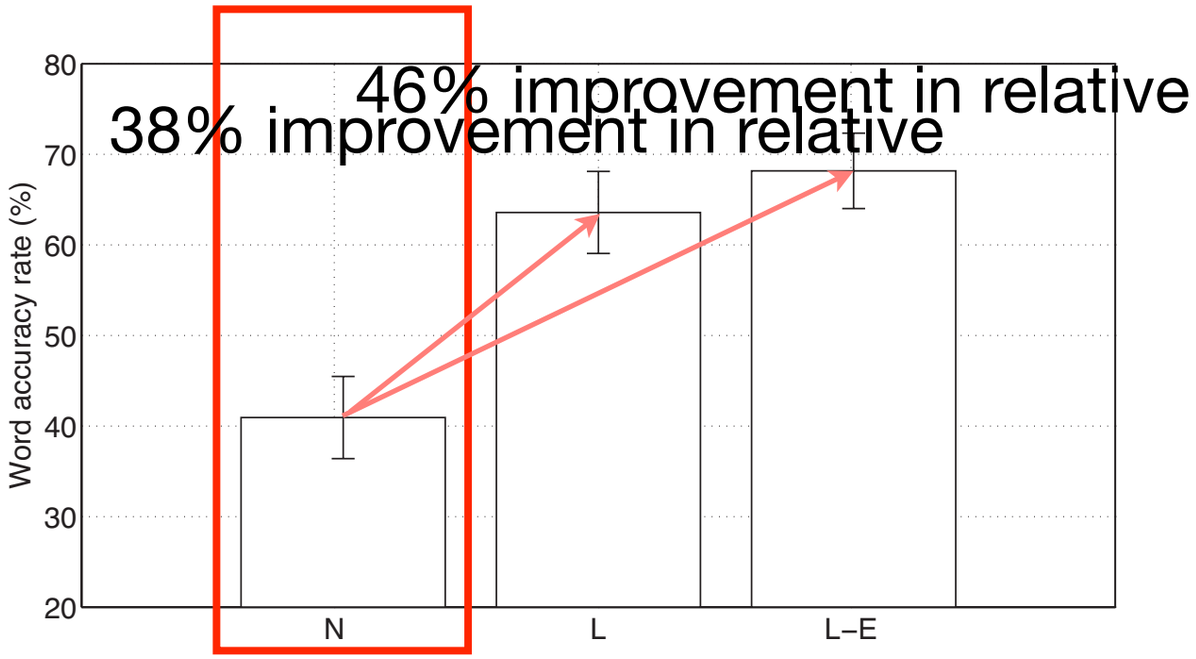
SEVENTH FRAMEWORK PROGRAMME

英語しか喋れない人から日本語の音声合成を作る！

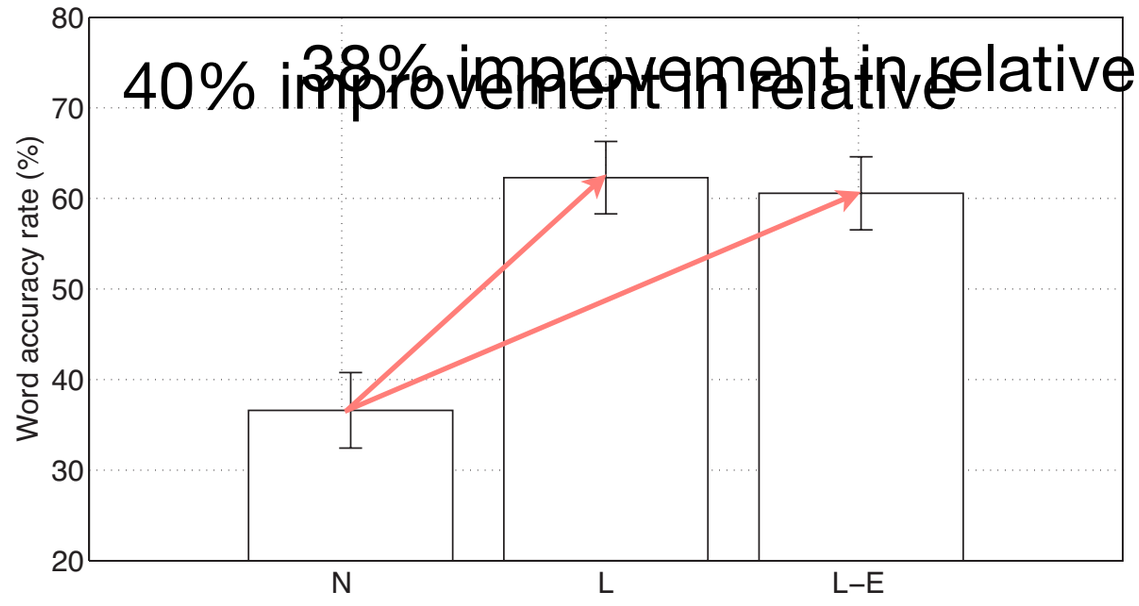
Target speaker ←



ボリュームを上げないでも騒音下で聞き易くなります



Speech modulated noise



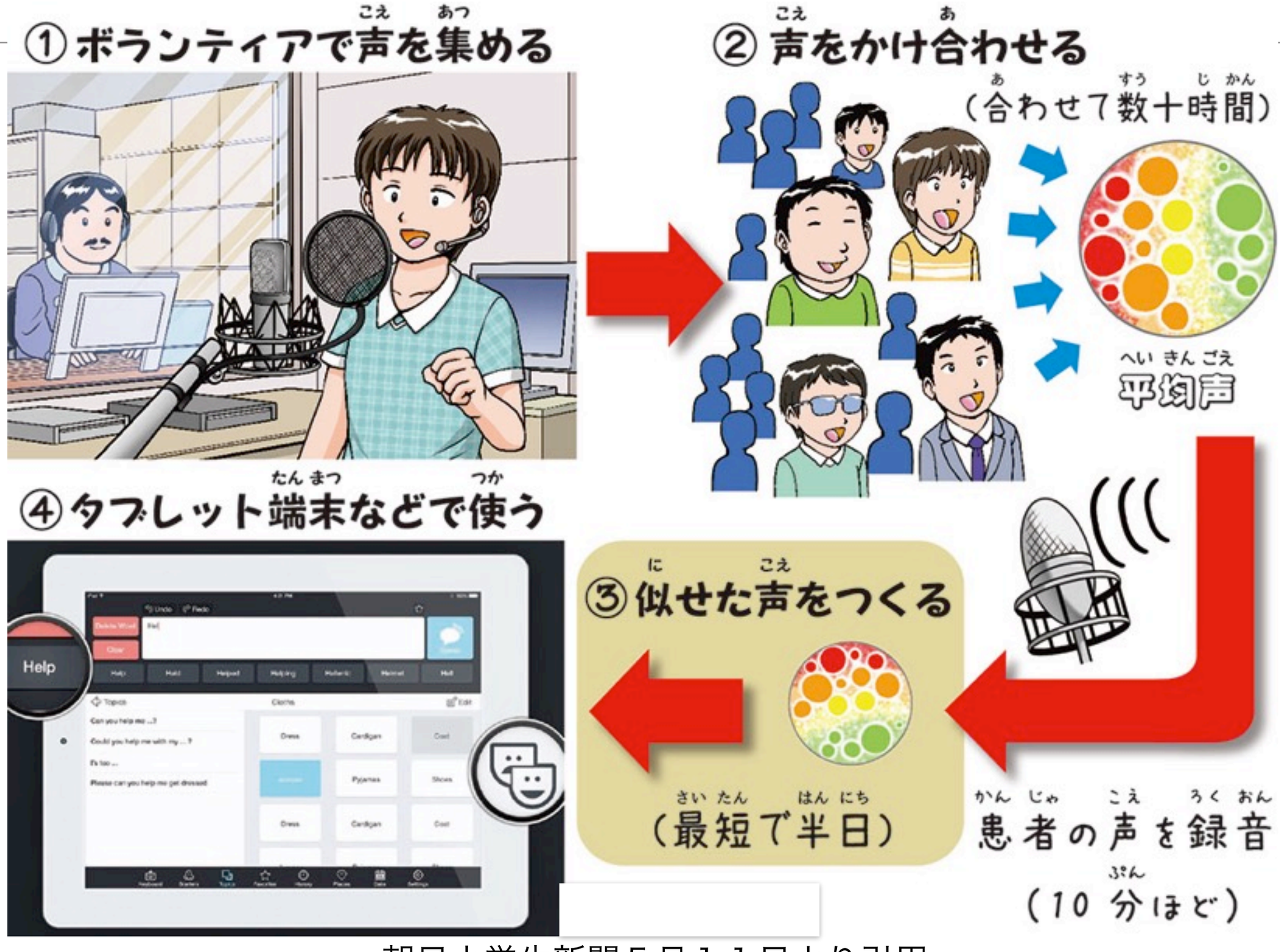
Competing speaker

C. Valentini-Botinhao, J. Yamagishi, S. King, “Mel cepstral coefficient modification based on the Glimpse Proportion measure for improving the intelligibility of HMM-generated synthetic speech in noise”, Proc Interspeech 2012

C. Valentini-Botinhaoa, J. Yamagishia, S. Kinga, R. Maiab ,"Intelligibility enhancement of HMM-generated speech in additive noise by modifying Mel cepstral coefficients to increase the Glimpse Proportion" Computer & Speech Language, 2012

声の障碍：ALS患者の場合

声に障碍がある人を助ける音声合成技術



朝日小学生新聞５月１１日より引用

まとめ

- 音声合成と最先端の統計的音声合成技術について紹介
- 合成音声の品質はかなり向上
- 音声の品質を良くする研究だけでなく、様々な応用技術の研究も実施
    - 自分の声で喋る音声翻訳システム
    - 騒音下でも聞きやすい音声合成システム
    - 声の障碍のある方の個人用音声合成システム

- 音声情報処理　面白いですよね？

丸善ライブラリー 「おしゃべりなコンピュータ」販売中！