自分の声でコミュニケーション！～音声合成技術最前線

山岸順一

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

自己紹介

- 音声情報処理、特に音声合成の研究に１４年間従事

- 2006年：博士号（東工大、工学）
- 2007〜現在：エジンバラ大 Senior Research Fellow
- 2013〜現在：国立情報学研究所 准教授

- 最近頂いた賞
  - 日本音響学会 独創研究奨励賞板倉記念
  - 情報処理学会 喜安記念業績賞
  - IEEE Signal Processing Society Young Author Best Paper Award
  - 文部科学省大臣表彰 若手科学者賞

本日の講演内容

- 音声合成
- 統計的音声合成
- 声の模倣、クローン技術
- 応用技術
    - 自分の声で外国語を喋る音声翻訳システム
    - 騒音下でも聞き取りや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
- 光と陰と未来

- 但し書き：ほとんどの実験が英国で行われたため、英語のデモが多いです

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, 喋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に変換
- 様々な応用例

音声合成の歴史１

Bell’s Voder
(1939)

Vocoder
(late 50’s)

音声合成の歴史２

Unit-selection synthesis
(mid 90’s)

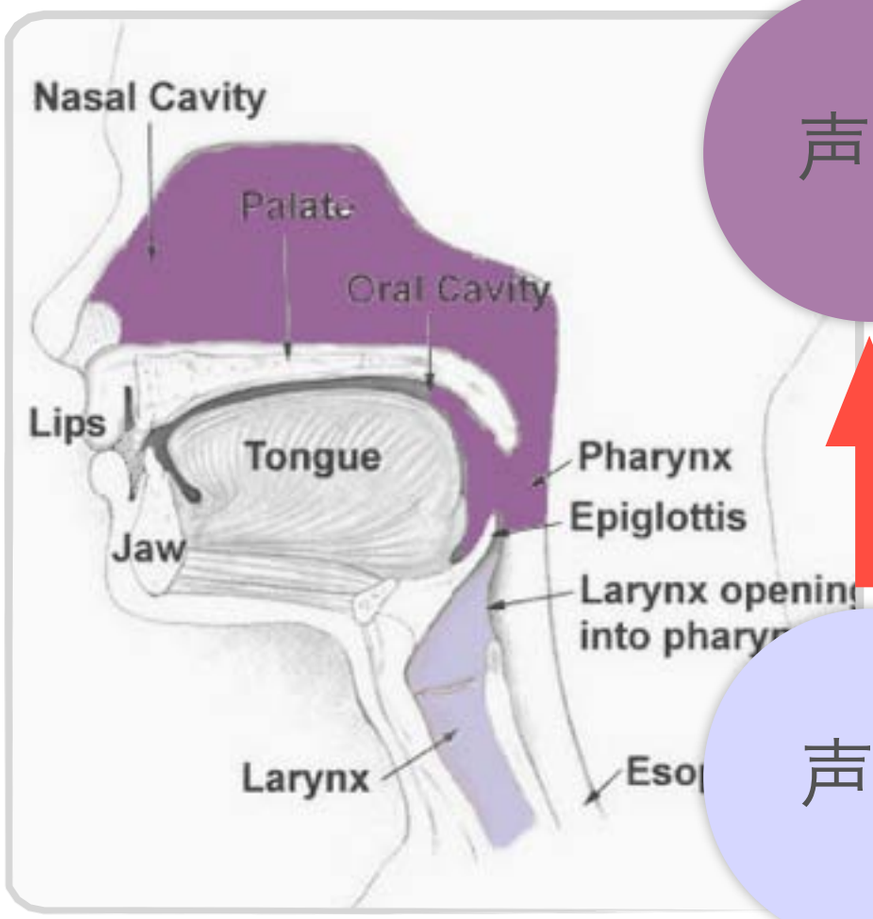
Diphone synthesis
(mid 80’s)

Statistical synthesis
(2000’s)

音声のパラメトリック表現と統計的音声合成

speech

Vocoder model



声道

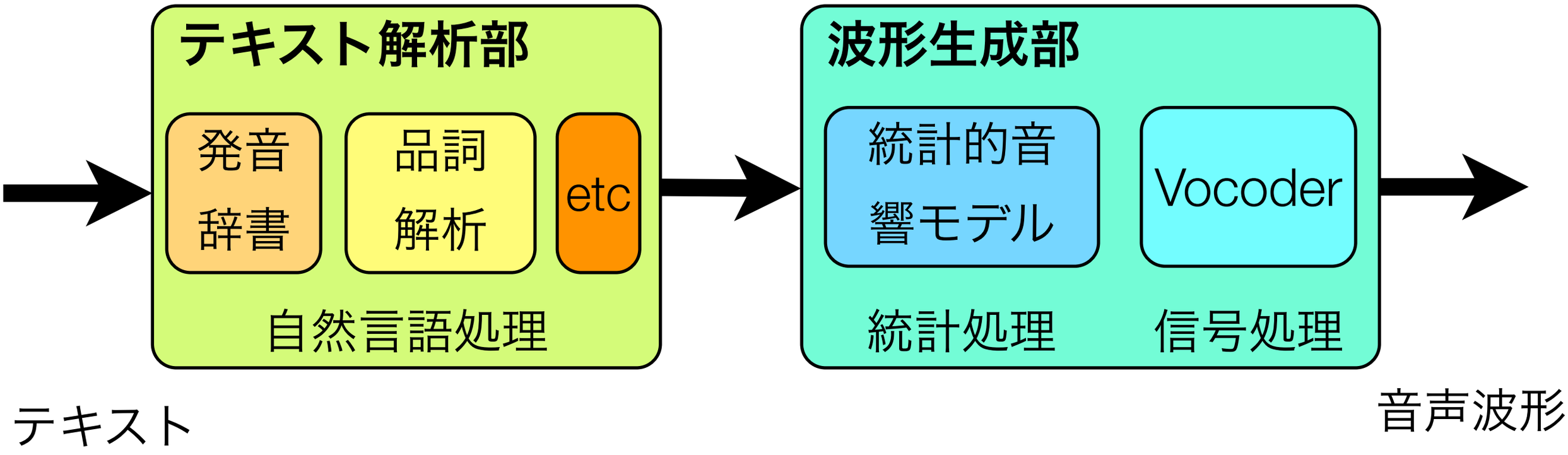
- Tone
- Articulation

声帯

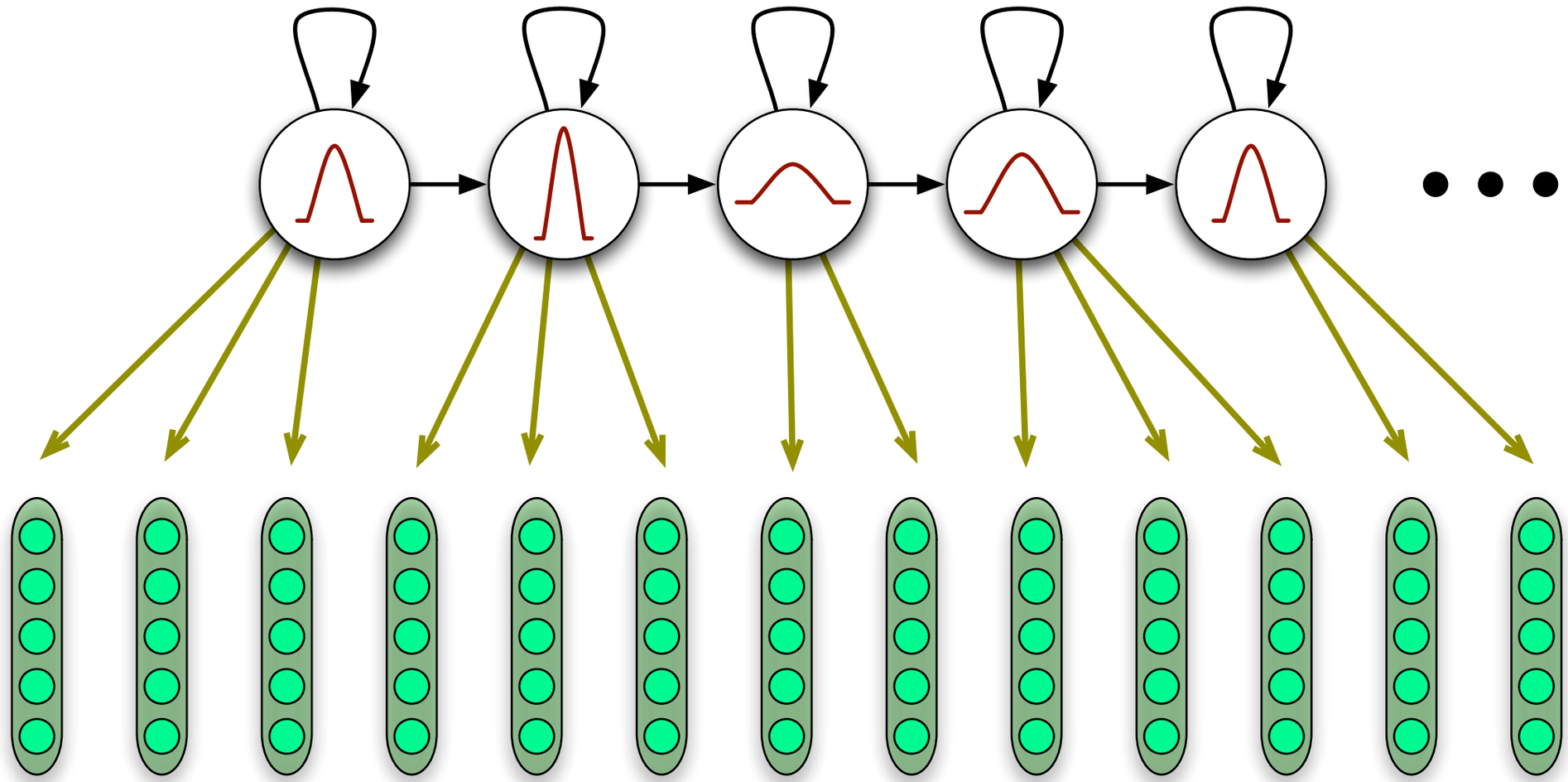
- Vocal quality
- Prosody

Statistical model

統計的音声合成システムの構成

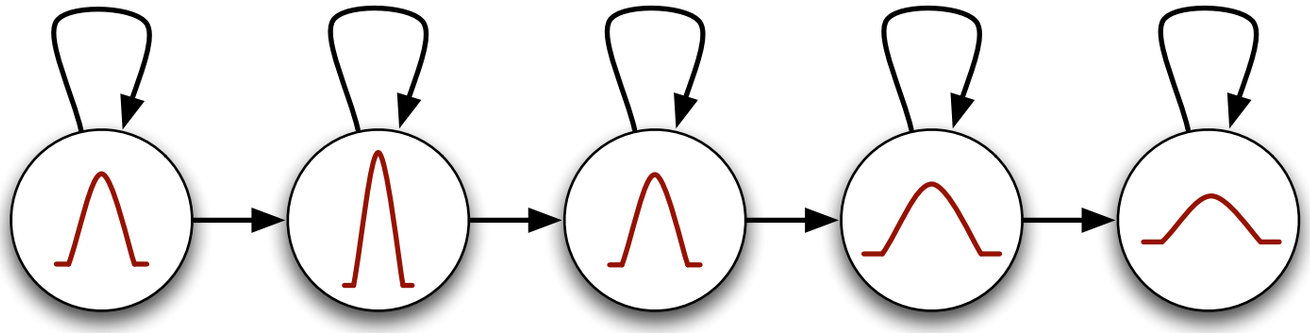


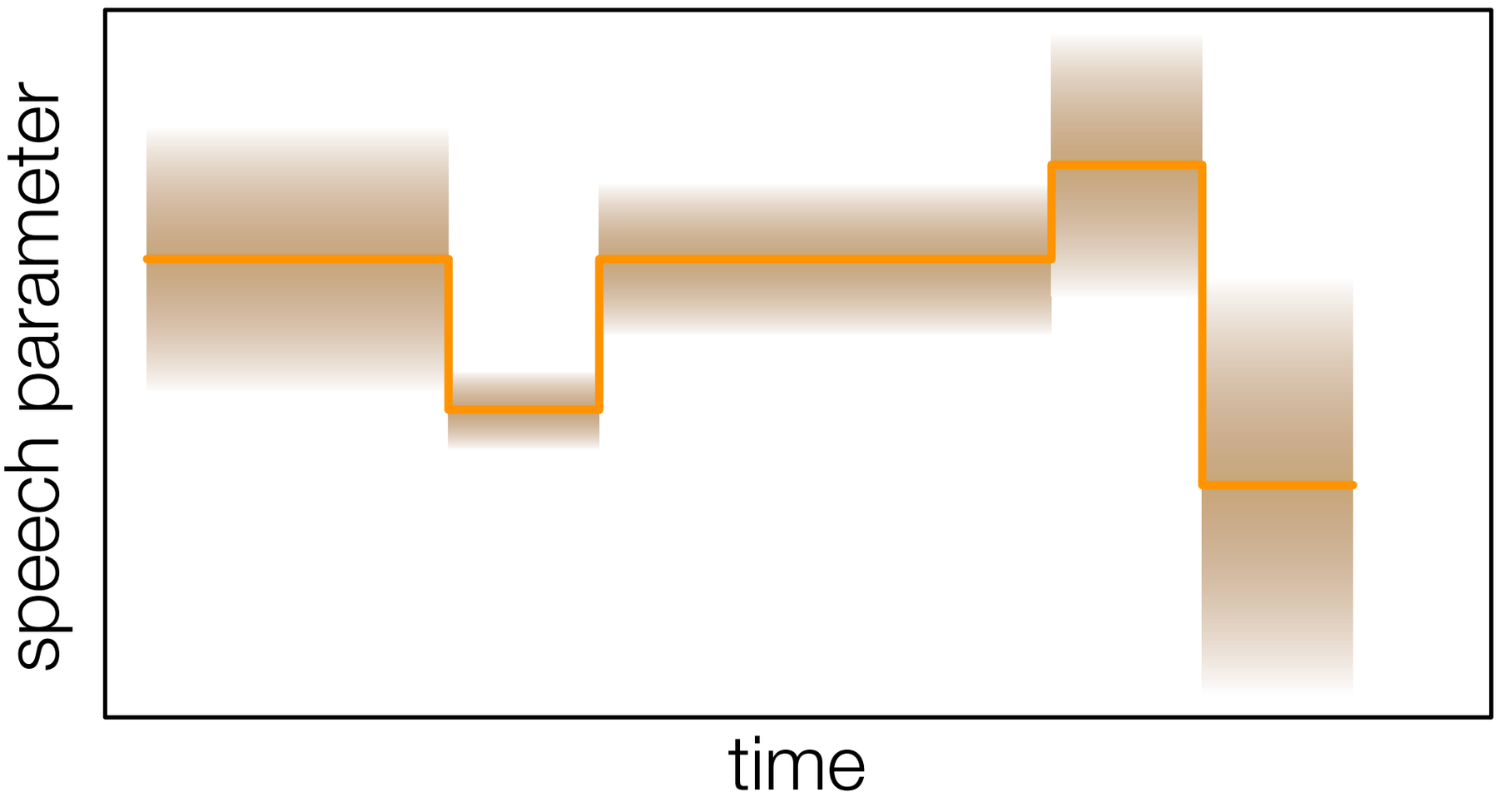
統計モデル：隠れマルコフモデル



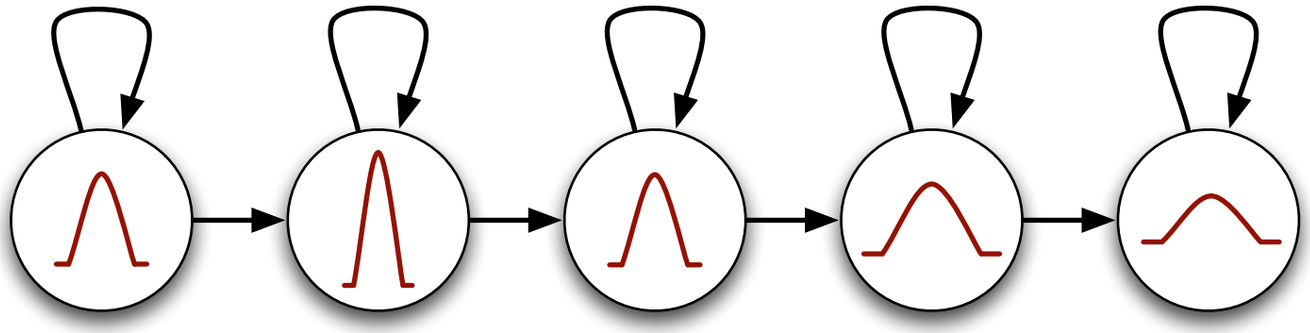
例：動詞の中の、前の音が子音「s」の母音「u」の音響モデル

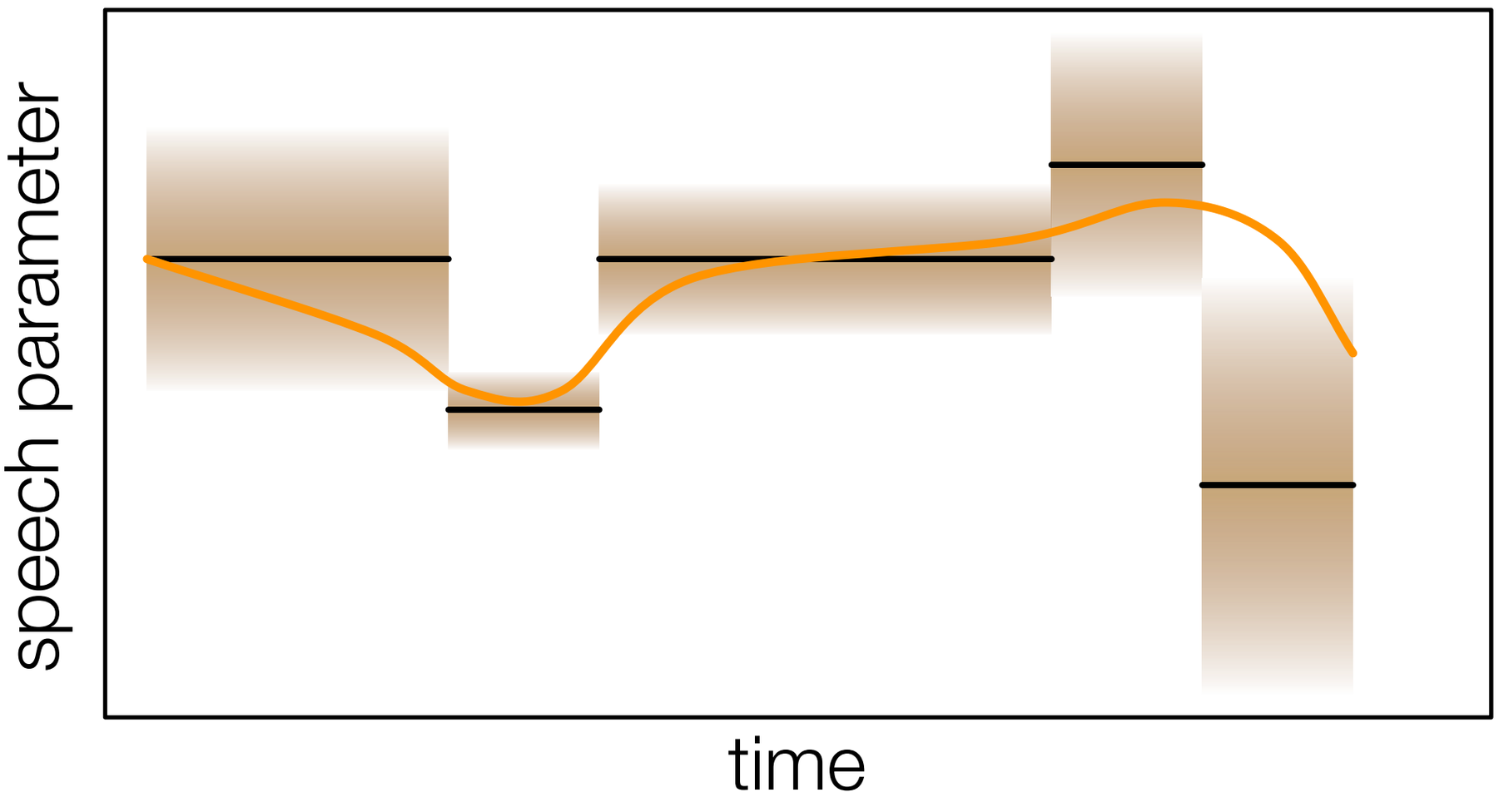
統計モデルからの音声パラメータの生成





統計モデルからの音声パラメータの生成





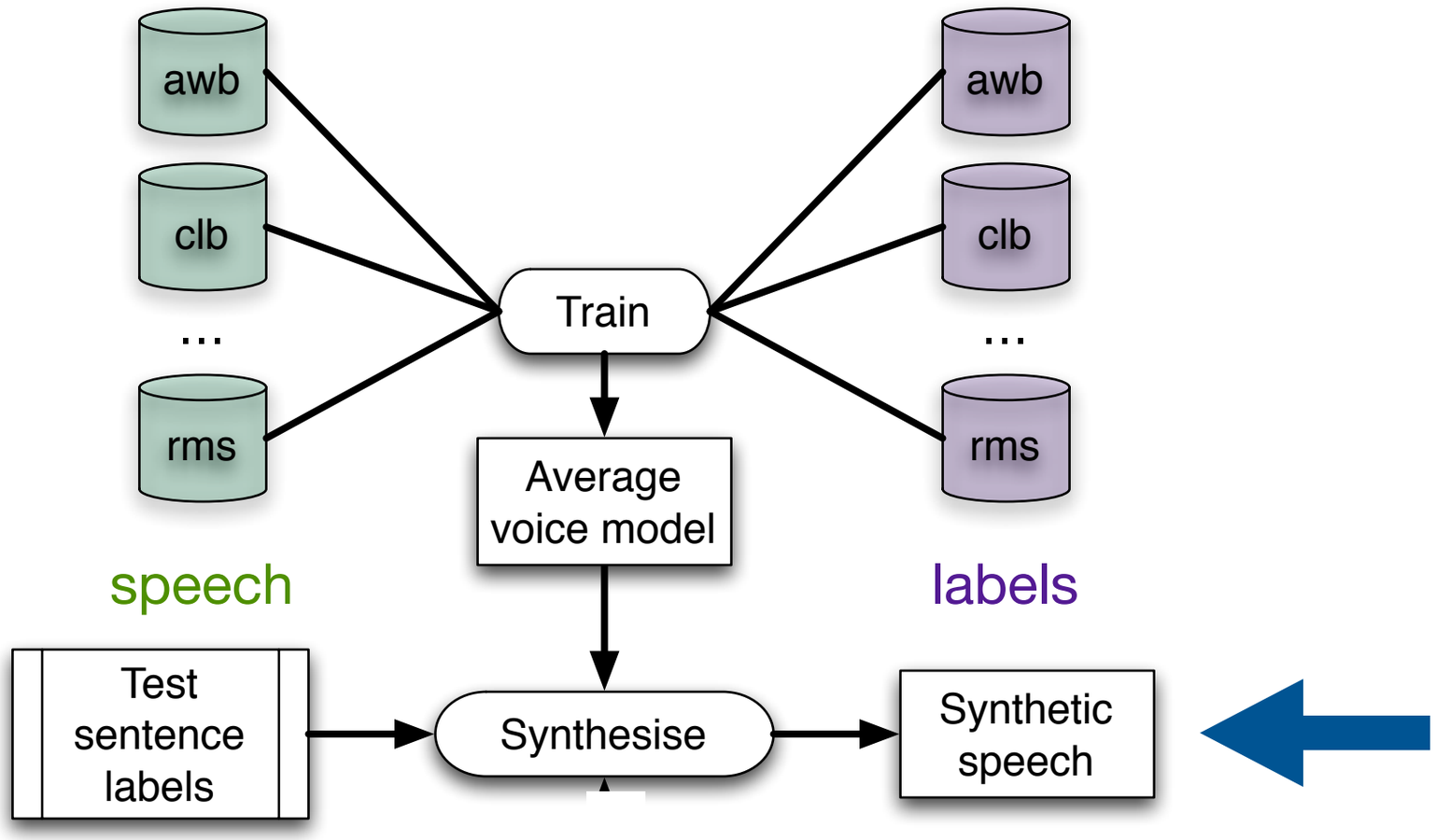
HMMに基づく音声合成

- 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〜)
  - 最新版の音声サンプル
  - 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

統計モデルを利用すると何が嬉しいのか

- 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

平均声のデモ：英語


awb
clb
...
rms
awb
clb
...
rms
Train
Average
voice model
speech
labels
Test
sentence
labels
Synthesise
Synthetic
speech

平均声のデモ：日本語

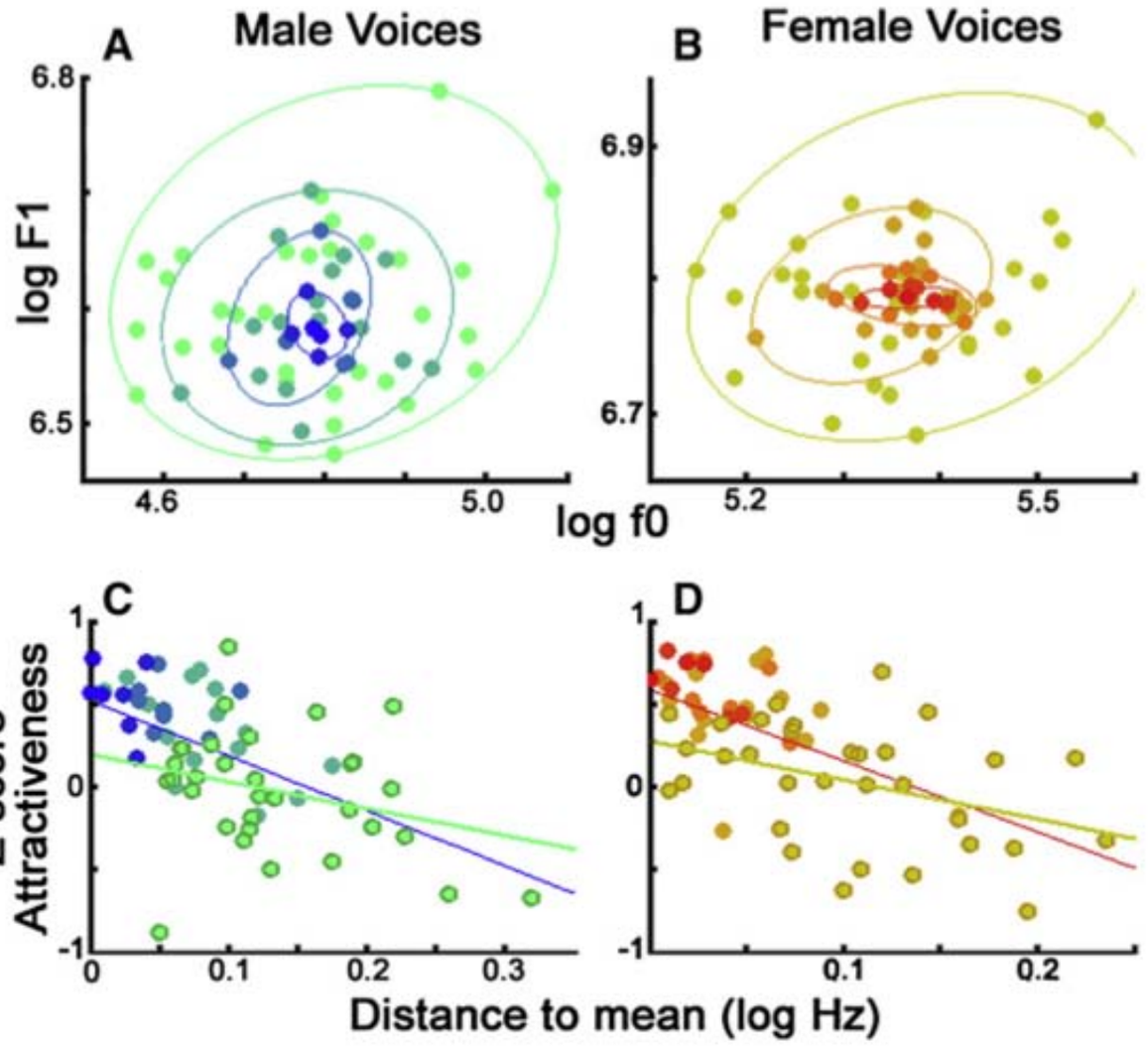
- いろいろな平均声が作れます
- 性別非依存平均声
    - 男性３０名
    - 女性３０名
    - サンプル

- 男性の平均声

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.

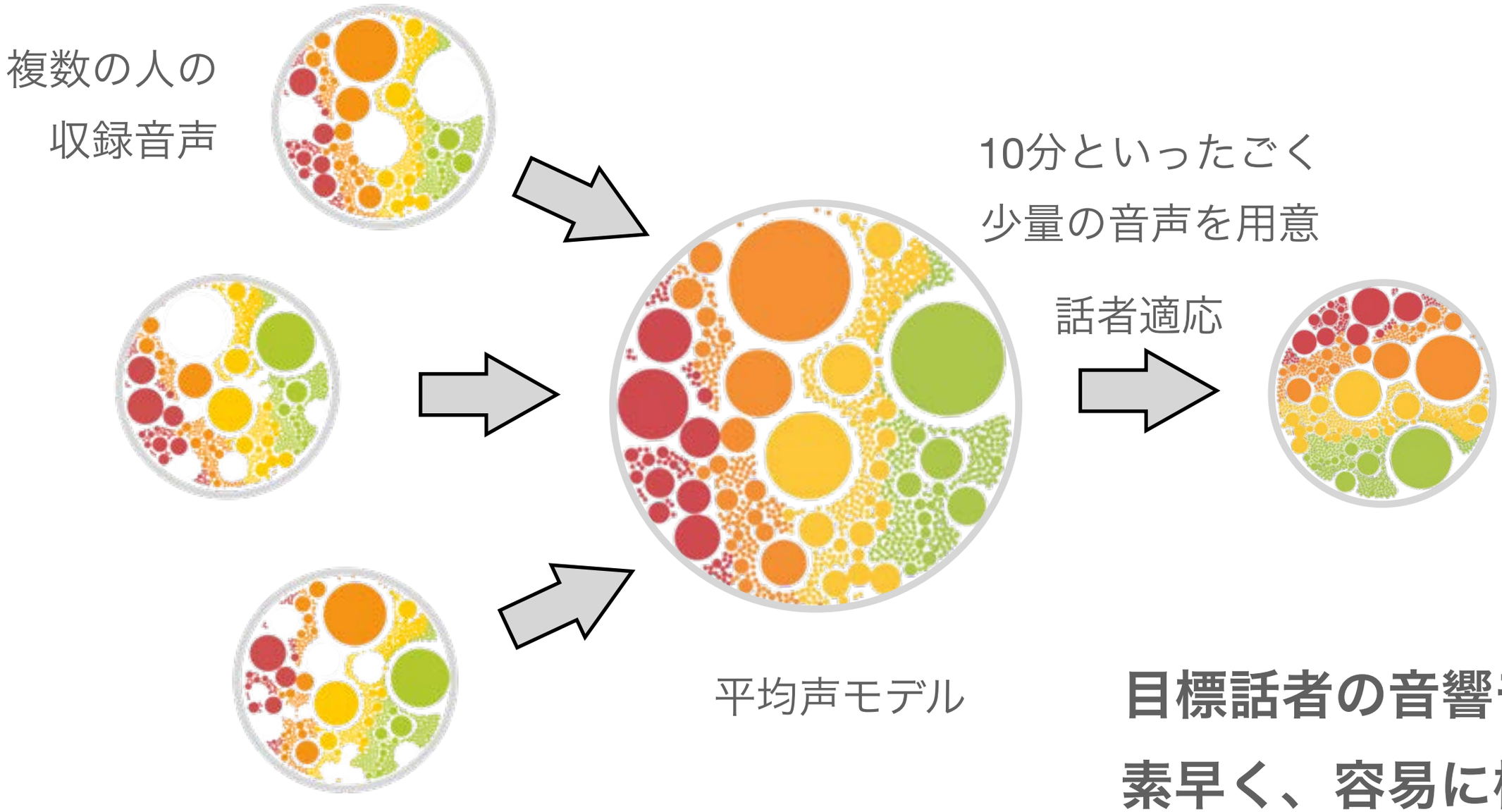
- 2、4、16、32名の平均声
- 基本周波数と第一フォルマント空間での平均声からの距離
- 声の魅力さのスコアとの相関係数：r= -0.59



平均声を利用すると、工学的に何が嬉しいのか

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

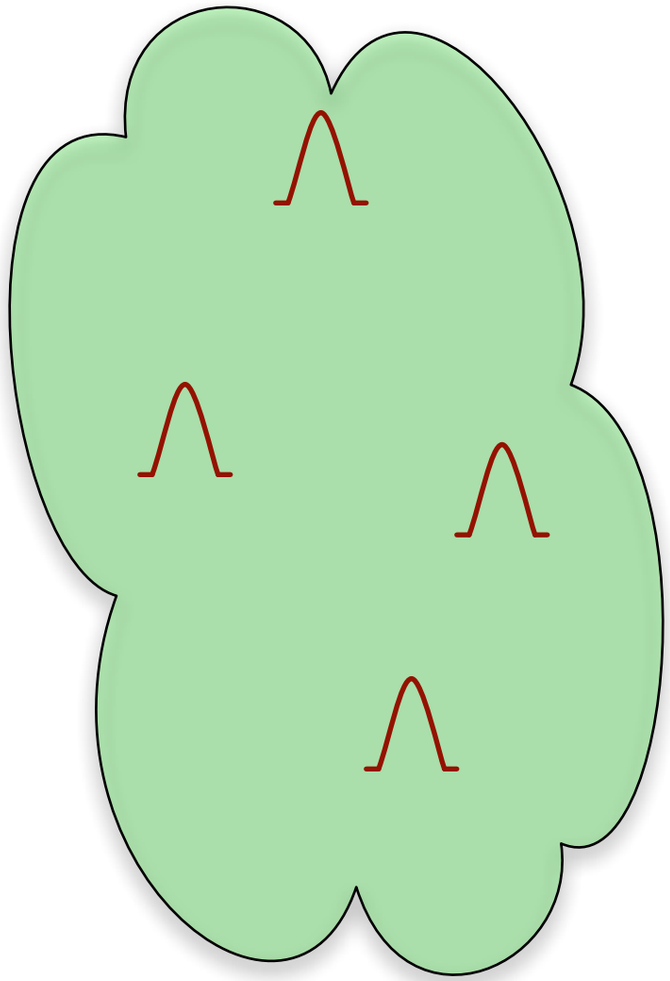
平均声を利用すると、工学的に何が嬉しいのか

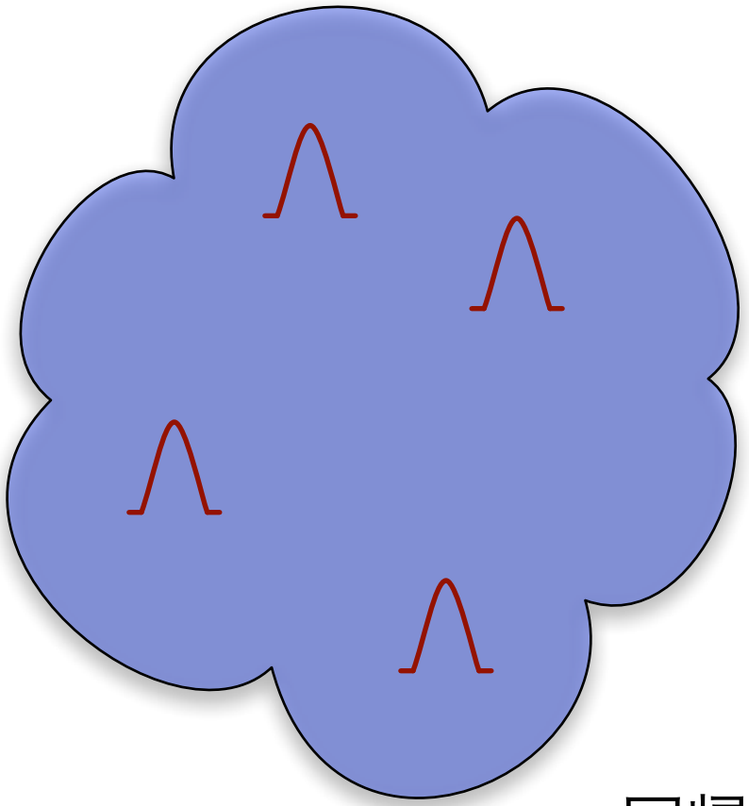


目標話者の音響モデルを
素早く、容易に構築可能

話者適応：統計モデルのアフィン変換

回帰クラス 1

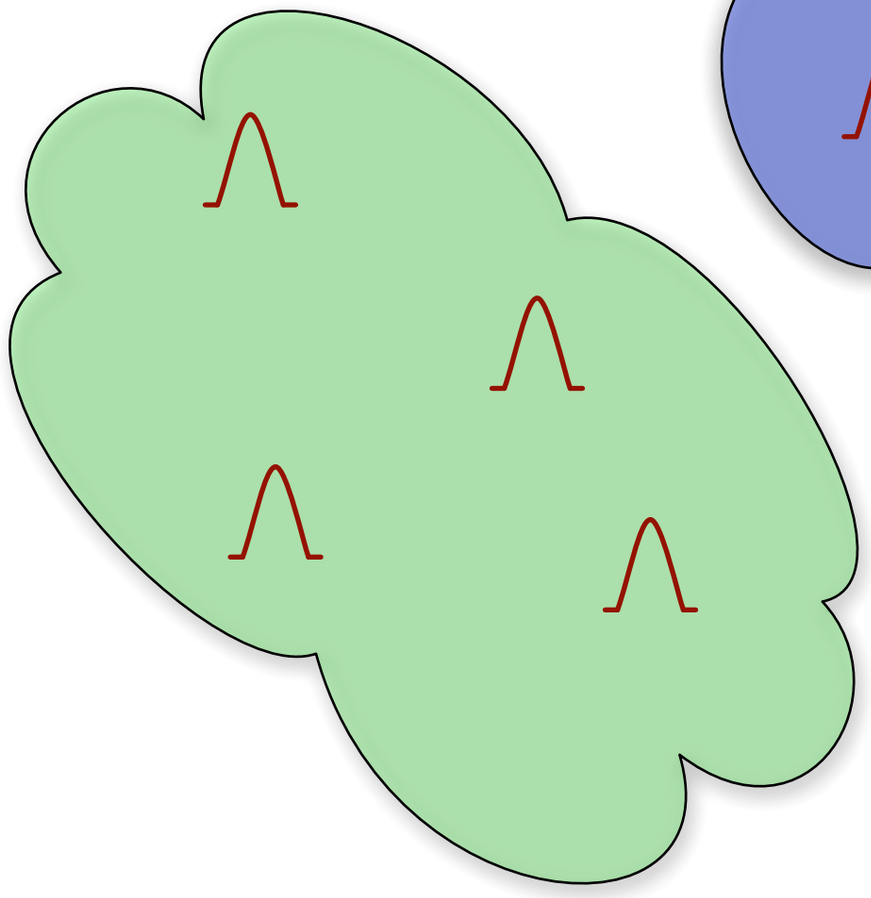


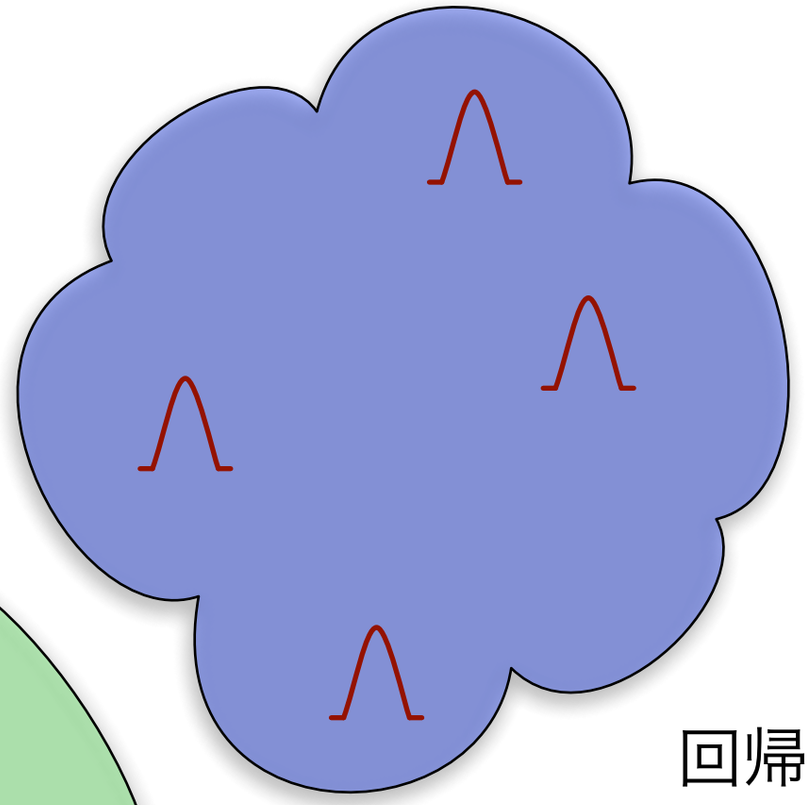


回帰クラス 2

話者適応：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

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

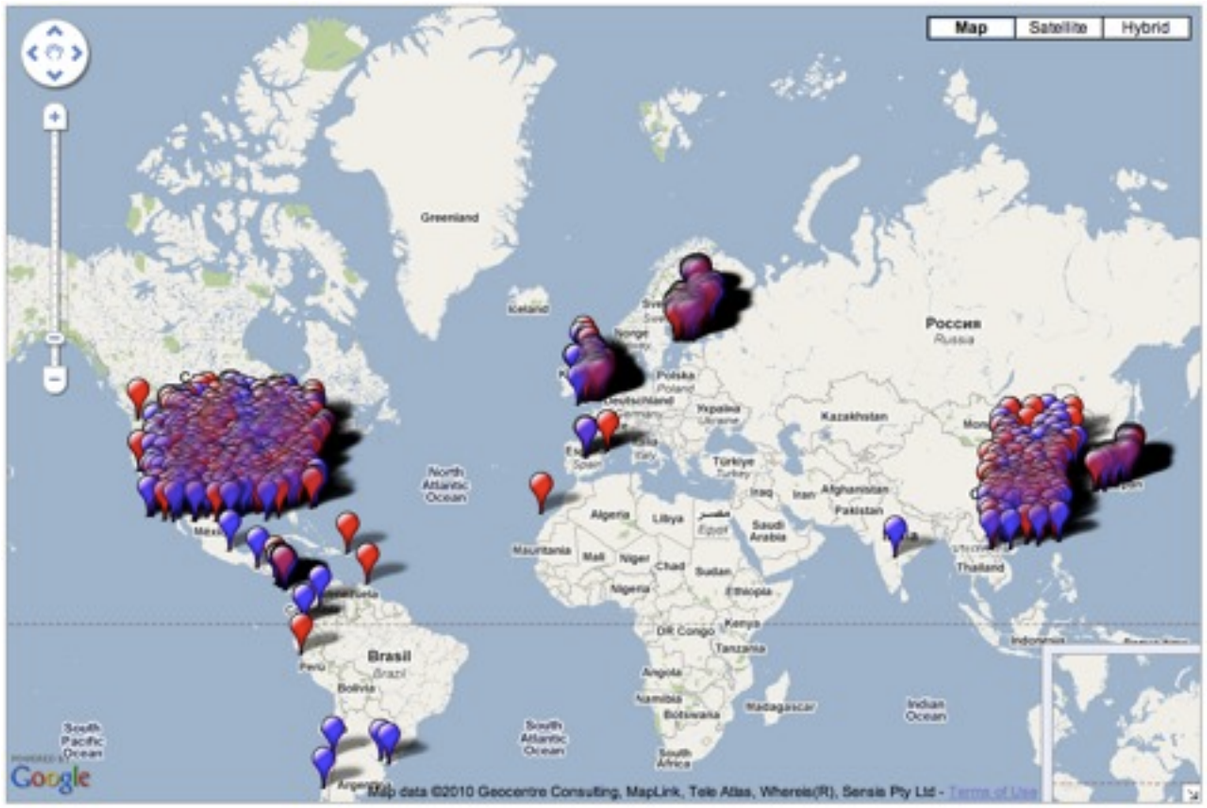
- アメリカ男性平均声からインド英語への適応
  - 平均声
  - 適応結果

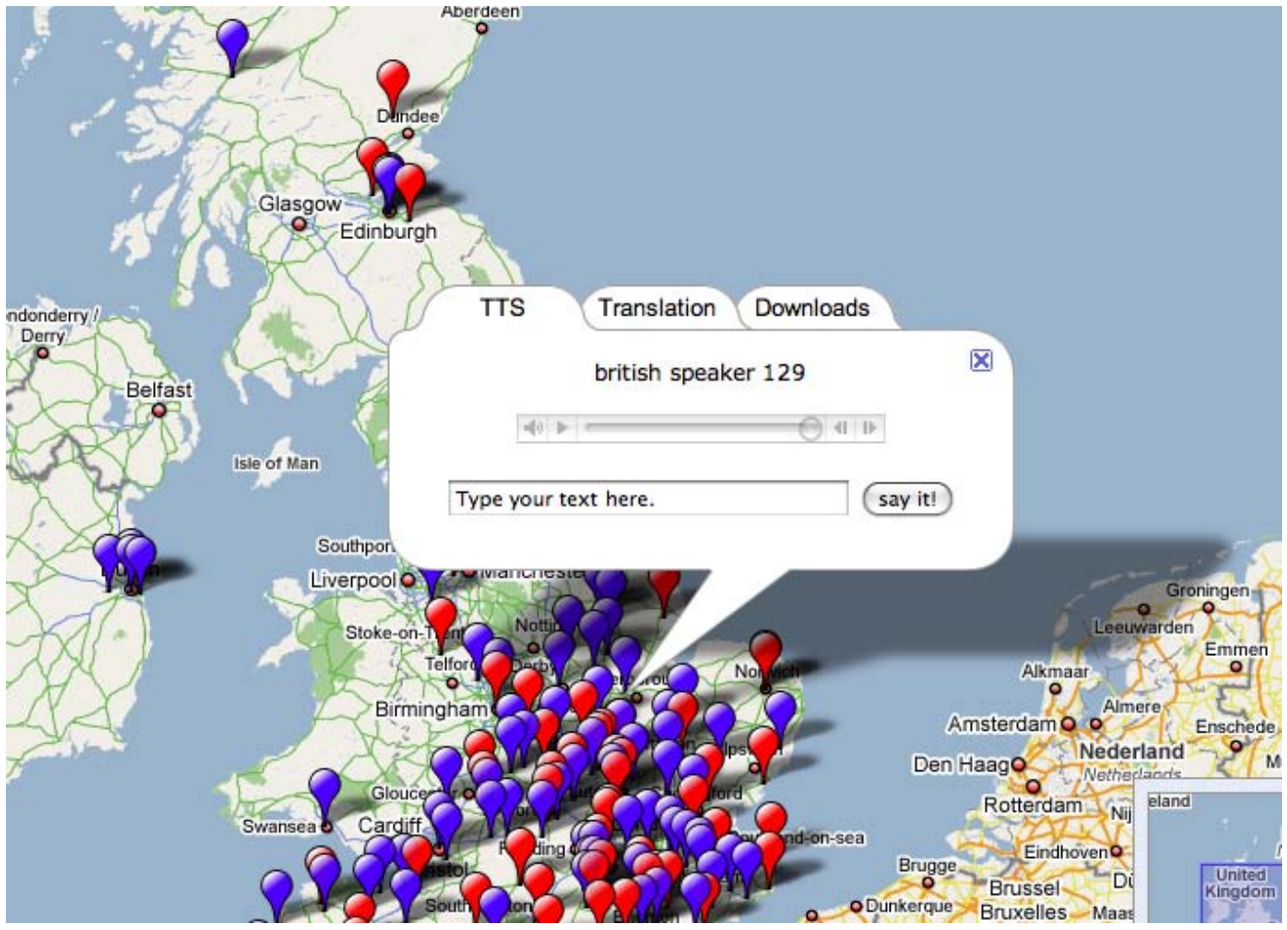
- 平静から怒りへの適応

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

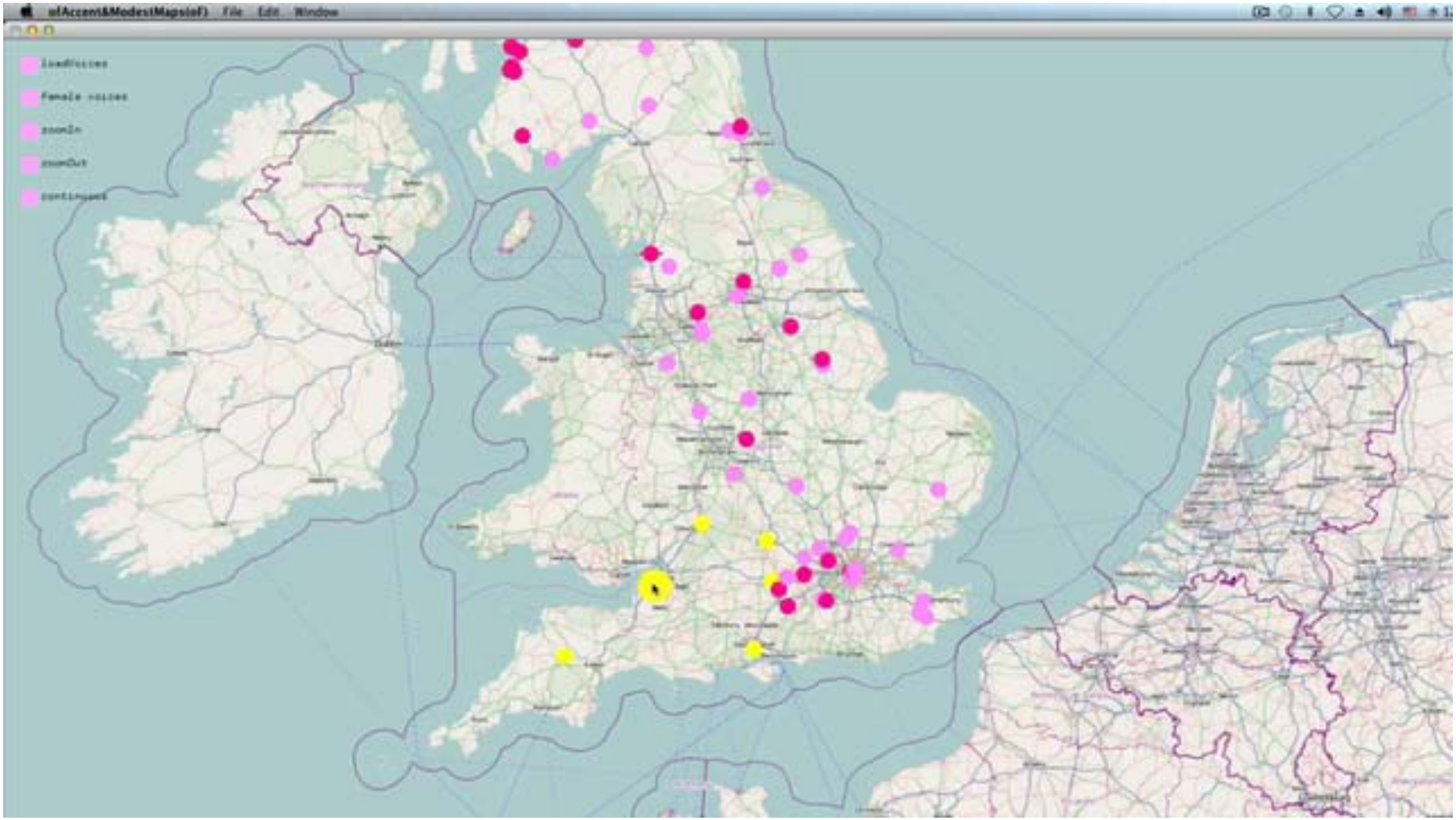




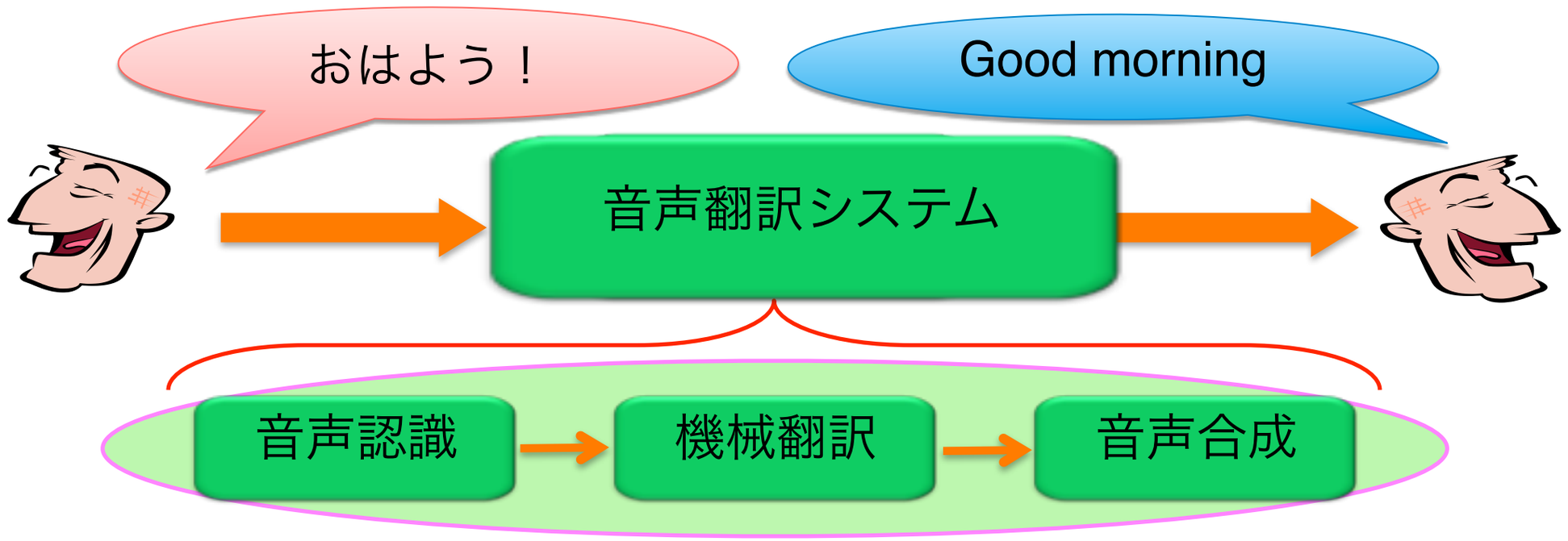
http://www.emime.org

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

各地の方言を音声合成で気軽に聞いてみたり、、、



自分の声で音声翻訳！

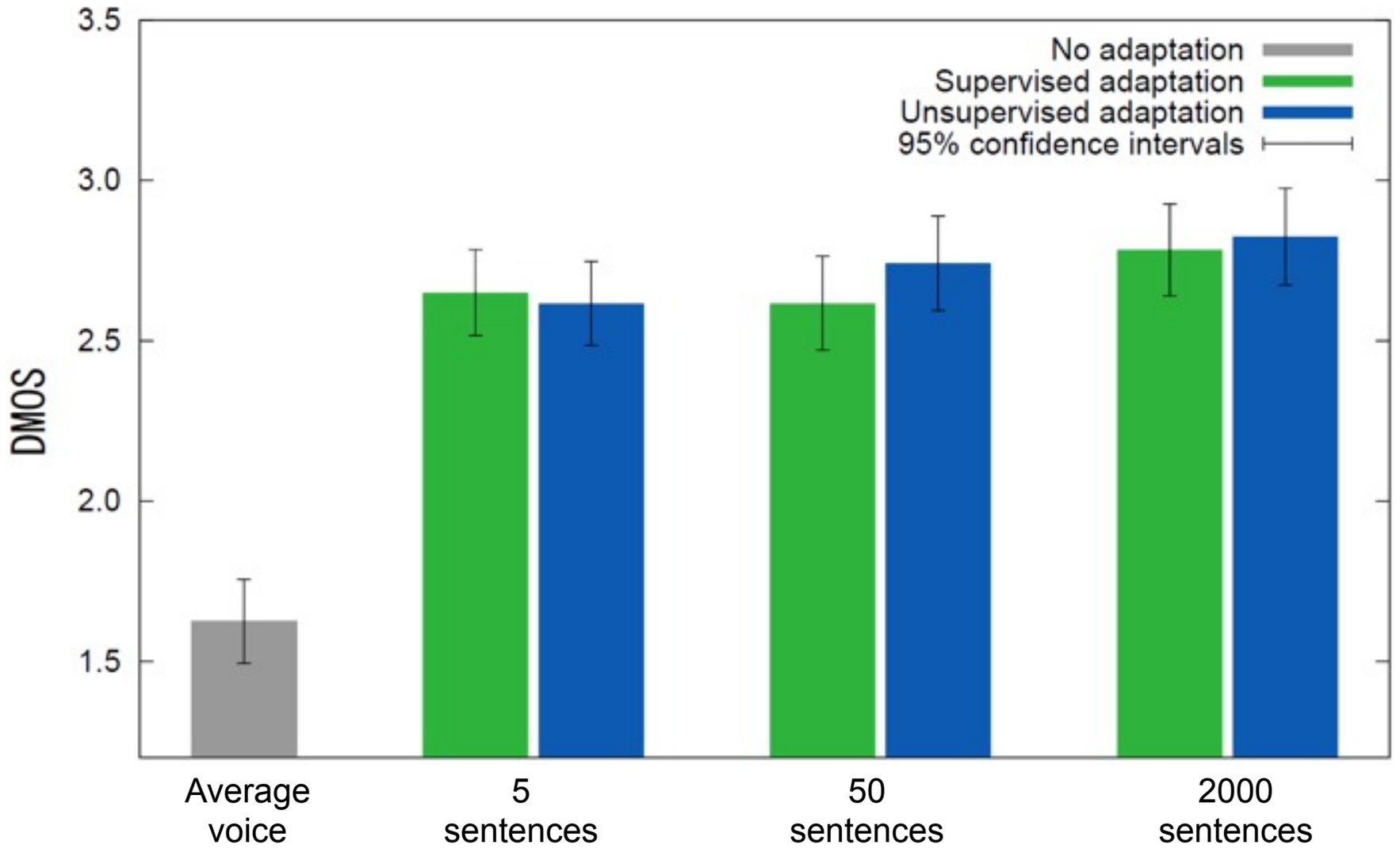


EMIME
Effective Multilingual Interaction in Mobile Environments

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



”耳”をもった音声合成：ロンバード効果の利用

Table 1: *A* *L-E voices.*

N		B/octave)
T		
Lombar		

LISTA

SEVENTH FRAMEWORK PROGRAMME

”耳”をもった音声合成：ロンバード効果の利用

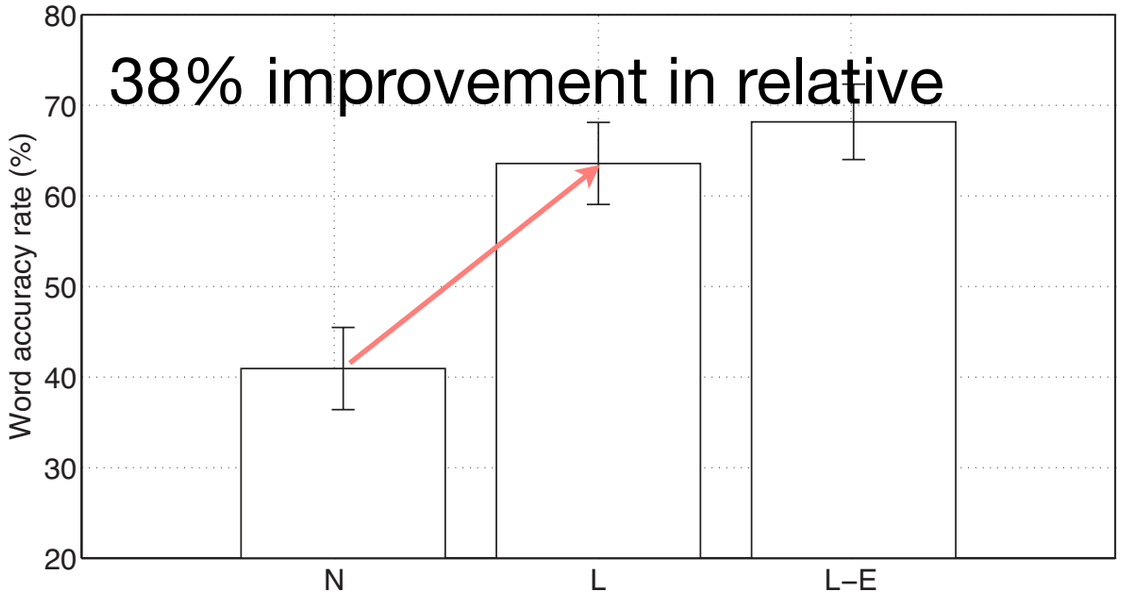
Table 1: *Acoustic properties observed in normal N, Lombard L, and extrapolated Lombard L-E voices.*

Voice	speech (secs.)	pauses (secs.)	F0 (Hz)	spectral tilt (dB/octave)
Natural speech				
Normal	2.06	-	107.1	-2.02
Lombard	2.32	-	136.8	-1.73
Text-to-speech				
Normal (N)	2.11	0.16	104.5	-2.09
Lombard (L)	2.80	0.19	145.0	-1.59
Lombard extrapolated (L-E)	3.05	0.20	144.8	-1.50

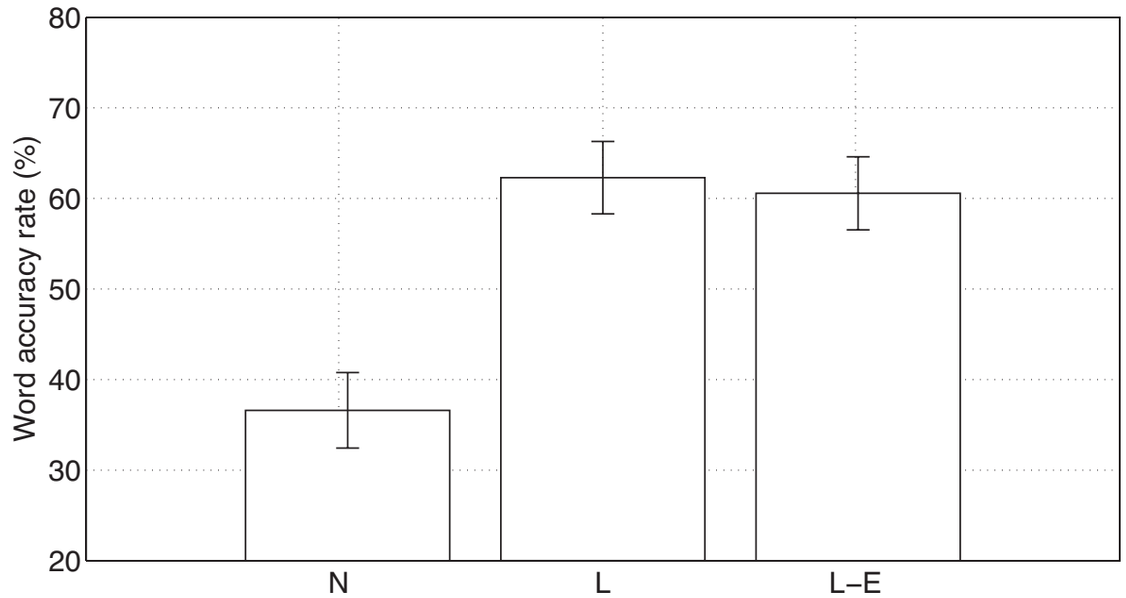
LISTA

SEVENTH FRAMEWORK PROGRAMME

ボリュームを上げないでも騒音下で聞きやすなります



Speech modulated noise

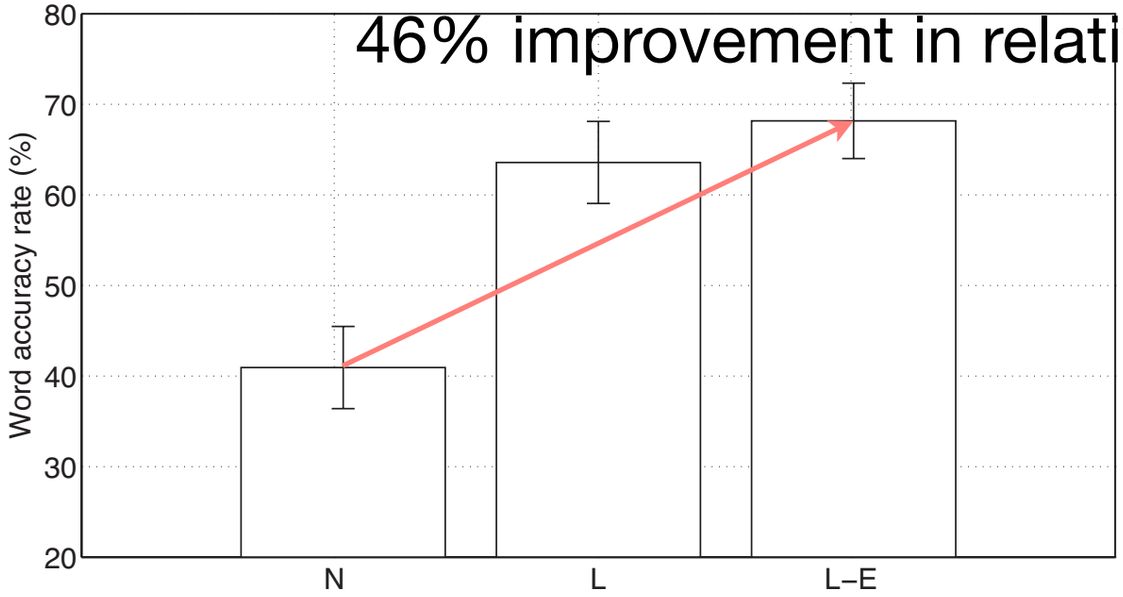


Competing speaker

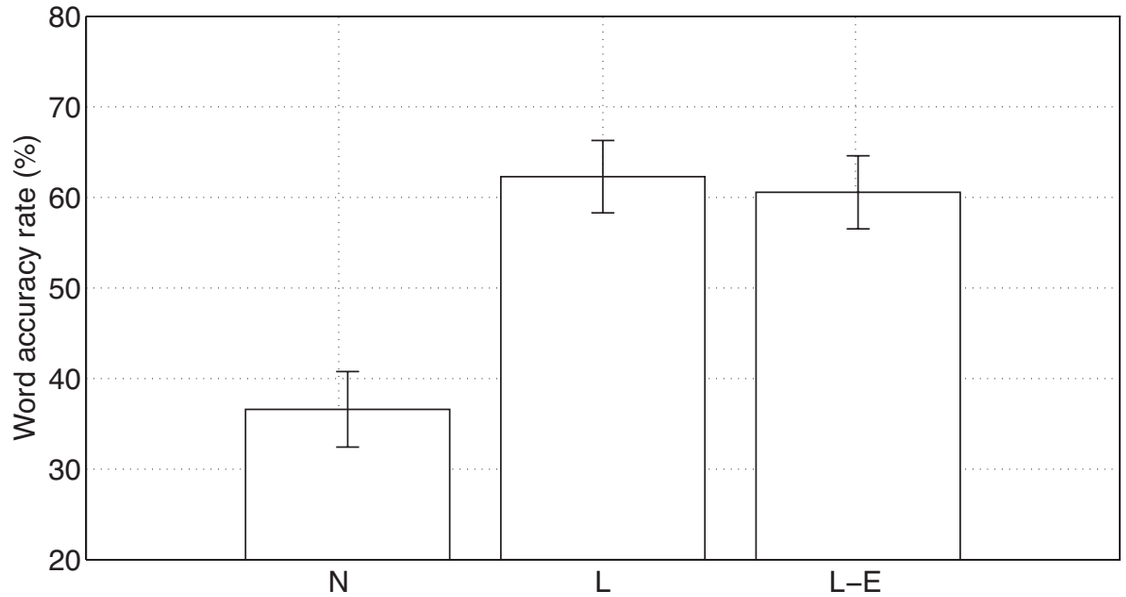
C. Valentini-Botinhao, J. Yamagishi, S. King, “Mel cepstral coefficient modification based on the Glimpse Proportion measure for improving the intelligibility of HMM-generated synthetic speech in noise”, Proc Interspeech 2012

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

ボリュームを上げないでも騒音下で聞きやすなります



Speech modulated noise

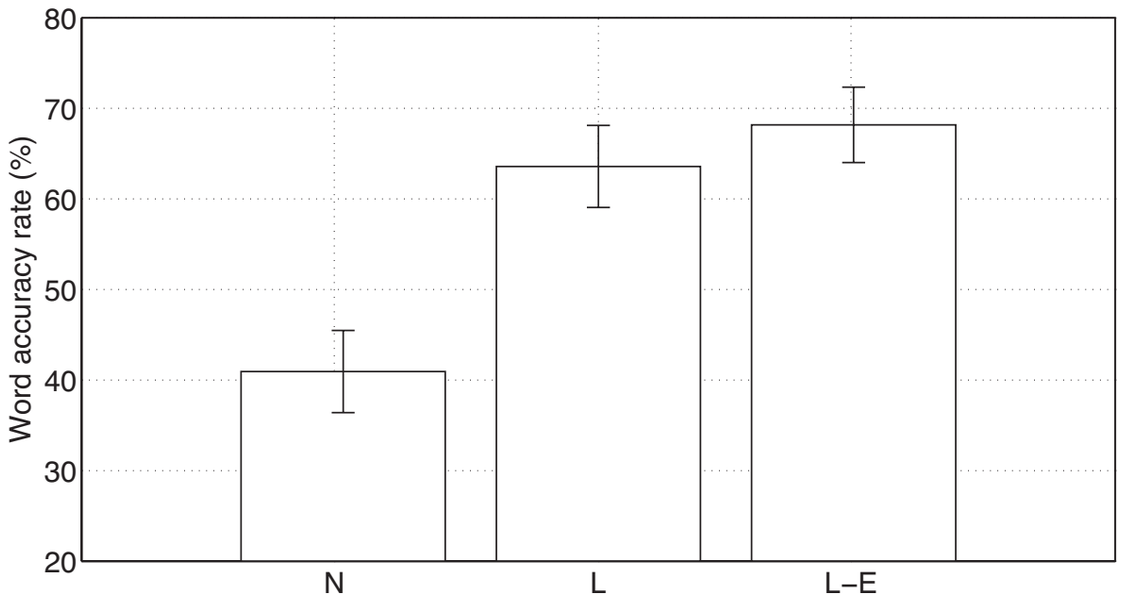


Competing speaker

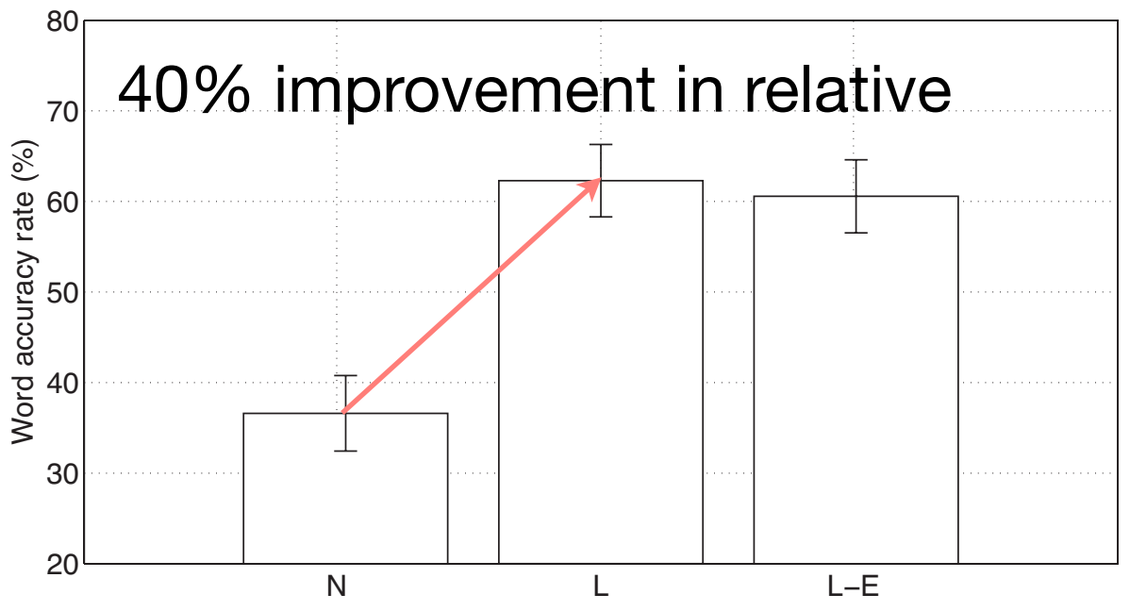
C. Valentini-Botinhao, J. Yamagishi, S. King, “Mel cepstral coefficient modification based on the Glimpse Proportion measure for improving the intelligibility of HMM-generated synthetic speech in noise”, Proc Interspeech 2012

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

ボリュームを上げないでも騒音下で聞きやすなります



Speech modulated noise

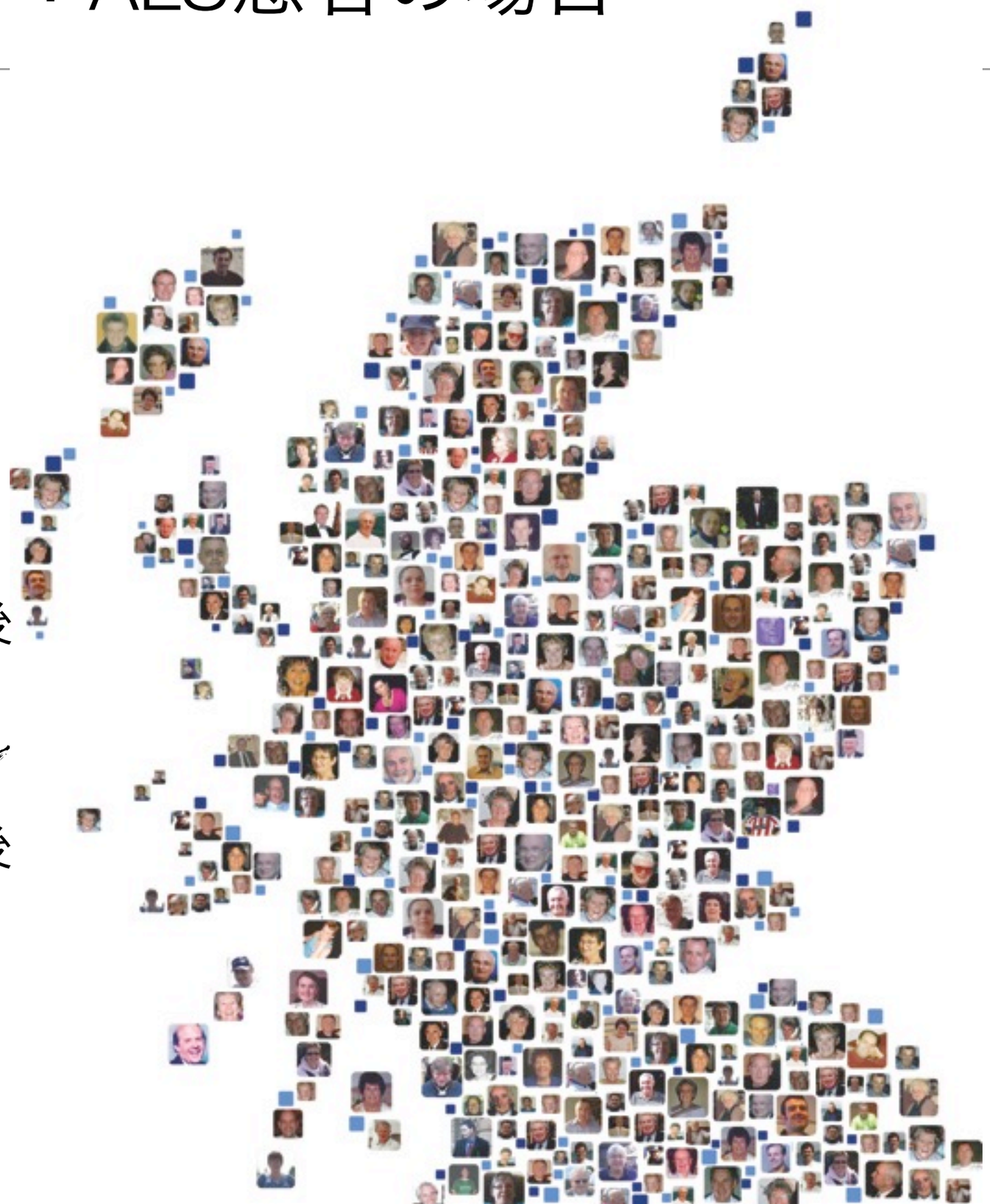


Competing speaker

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



診断直後

↓

８ヶ月後

診断直後

↓

9ヶ月後

意思伝達装置

意思伝達装置は音声出力も可能だが、、、

- 現状だと、「声」の選択肢はあまり無い
  - 通常１つか２つ程
- 年齢、方言、発話様式を音声出力に適切に反映できているとは言いがたい
- 現在、１、２社が個人の声から音声合成システムを構築するサービスを提供しているが、非常に高価
  - 約１００万

- 声はコミュニケーションの手段のみならず、アイデンティティでもある
- 個人の声による音声合成システムの普及は、意思伝達装置ユーザに非常に求められている

Michael Arnott

second individual to receive a personalised voice

ALS患者の自分の声を再現する音声合成システム

- 英国ユアンマクドナルドMND研究所との共同実験
- MND (ALS)との診断直後に、音声を２０分収録
- 2011年の収録時には、構音障害無し
- 9ヶ月後、症状が進行し、構音障害が発声したため、会話補助アプリとして音声合成を届ける
- 同じ地域にすむ健常者の声を２０名集め、平均声を作成。話者適応を行う

ALS患者の自分の声を再現する音声合成システム

ボイスバンククラウドツール（一部開発中）

音声収録アプリ

アップロード

音声合成システム
自動構築クラウド

iOSアプリ

意思伝達装置対応

SAPI5フォーマット

VoiceClone

Now communication made easy.
Communicate by using an artificial voice of your choice.

Voice Bankプロジェクト

- 英国ユアンマクドナルドMND研究所とアニーローリング再生医療病院との共同プロジェクト
- ２０１１年～
- 英国７００人
- 日本４００人

声のボランティア

声の障碍に備え
自分の声を保存

音声データを提供

National Institute of Informatics

THE UNIVERSITY OF EDINBURGH

Voice Bank

自分の声を利用した
音声合成システム

音声

音声の障碍者

音声

自分の声を利用した
音声合成システム

音声の障碍者

“Vocal Terror” [BBC 2007]

BBC NEWS

Scientists warn of 'vocal terror'

By Liz Seward
Science reporter, York

Computers could mimic human speech so perfectly that vocal terrorism could be a new threat in 10-15 years' time, scientists suggest.

In the future, it may be possible to mimic someone's voice exactly after recording just one sentence.

Such technologies would pose a danger if it were not possible to verify who was speaking, researchers believe.

Scientists were predicting the future at the British Association (BA) Festival of Science in York.

Dr David Howard from the University of York said: "The reason things are changing is because no longer are we using an acoustic model proposed in the 1950s."

> “ **It's not scaremongering; it's trying to say to people, 'we have to think about these things'** ”
> David Howard

New methods of creating computerised speech use models of a vocal tract to create a realistic sound, replacing the existing technique of copying sounds.

声の認証システム vs 声のクローン

- 手軽な生体認証システムとして、声を使った認証システムがあります
- クローンの技術の光と陰
- 個人用音声合成システムは、声の認証システムを破れます
- ３００人分のデータを使った話者識別システム

- GMM-UBM/SVM GMM kernel/iVectorの標準的話者照合システム３種
- オープンソースで公開されている技術のみで、「声の詐称」を検証
- 声のクローン技術は、３００人中２８８人の詐称に成功

声の認証も声のクローンもどちらも安心して使える未来

- 合成音声を自然音声の聴覚上の差は小さくなっています。
- ゆくゆくは聴覚上ほぼ同じになるはず
- それでも自然音声と合成音声を区別する方法はあります。
- 例えば
  - 音の位相情報
  - 人間の耳は音の位相情報の違いに敏感ではない
  - しかし計算機はその違いを容易に検出できる
  - 300人の声を使った実験では、88%の確率で合成音声を正しく検出
  - その他の方法も研究されています

まとめ

- 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
- 声の認証システムも声のクローンもどちらも安心して使える様、合成された音声を検出するアルゴリズムについても検討

- 音声情報処理　面白いですよね？