賢い機械の口・耳を創る：
怒るコンピュータ，歌うコンピュータ，そして聞き耳をたてるコンピュータ

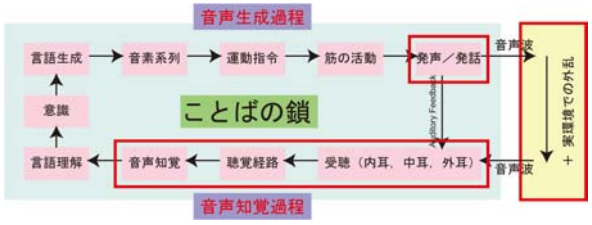
赤木　正人
教授　情報科学研究科
北陸先端科学技術大学院大学（JAIST）

JAIST
JAPAN ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY HOKURIKU 1990

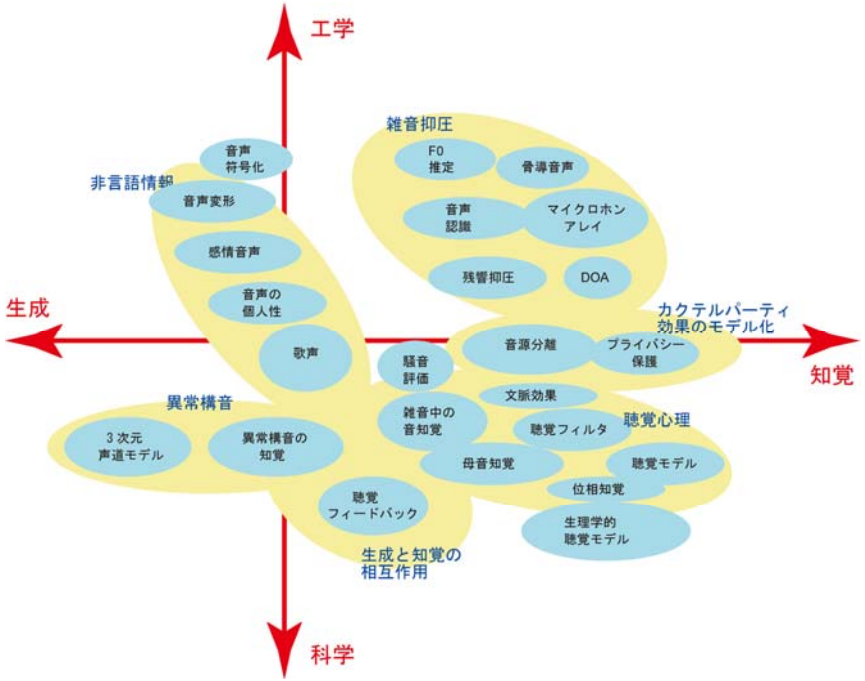
JAIST

基本路線

- **話す・聞く**は人間の営み　→　人間を知りそして営みを記述することで，高度の音処理システムの実現を目指す
- **研究範囲**：音声知覚，音声生成



ヒトの観測　→　モデル化　→　工学的応用



今日の話題

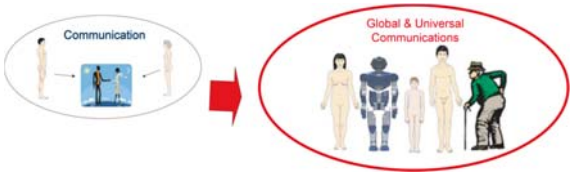
1. 怒るコンピュータ、歌うコンピュータ（非言語情報）
   - 音声の声質に関する知覚モデルの提案「ヒトは声の印象をどのように知覚しているか？」
   - モデルの応用「合成音声への非言語情報付加」：２つの例
     1. 感情音声の合成
     2. 歌声の合成
2. 聞き耳をたてるコンピュータ（カクテルパーティ効果のモデル化）
   - 聞き耳のルール「ヒトは多くの音の中から目的音をどのように知覚しているか？」
   - モデルの応用：２つの例
     1. 楽器音の分凝
     2. 音声のプライバシー保護

非言語情報の付加

怒るコンピュータ、歌うコンピュータ

ユニバーサルな音声コミュニケーション

- 言語・民族・文化を越えた（＝グローバルな），また，言語・民族・文化のみならず老人，幼児，あるいは障害者との障壁のない（＝ユニバーサルな）コミュニケーションの重要性が増している
  - 音声コミュニケーションでは，「何を話しているか」という言語情報だけではなく，これ以外の情報，たとえば，「誰が話しているのか」（個人性），「どんな気持ちで話しているのか」，「健康状態はどうか」などの非言語情報が多数送受される
- 言語情報＋非言語情報が一体となってコミュニケーションを円滑にする
  - 言語を越えたグローバルでユニバーサルな音声コミュニケーション環境の構築 ← 非言語情報についての音声コミュニケーションを解明することが一助となる

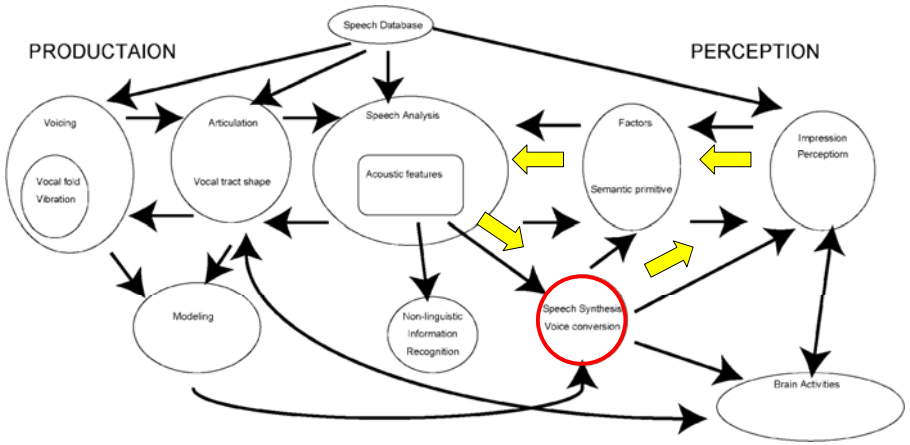


ユニバーサルな音声コミュニケーション(2)

- 言語・民族・文化を越えたユニバーサルコミュニケーションのためには
  - 非言語情報の生成・知覚において，言語・民族・文化によらないヒトの生物学的「共通要素」，すなわち，
    - 生成のための万国共通の構音運動，
    - 共通の構音運動から作り出される共通の音声特徴，
    - 音声特徴を呈示することにより生起される共通の知覚特徴・脳活動，そして，
    - この上に立つ人間の共通の行動

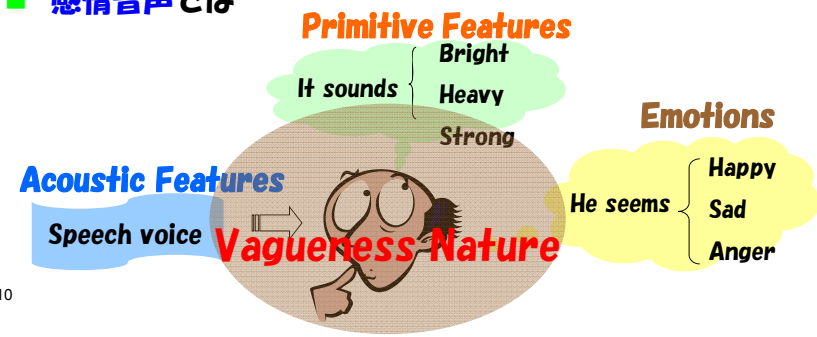
    が存在しなければならない．
- 研究方針
  - 音声の生成と知覚は不可分である．環構造（ことばの鎖）の中でのそれぞれの関係を考慮しながら研究を進める
  - 言語・民族・文化によらない非言語情報の生成・知覚機構の共通要素とは何かについて検討
  - 共通要素を核として，非言語情報の合成・認識を試みる

非言語情報：研究の流れ



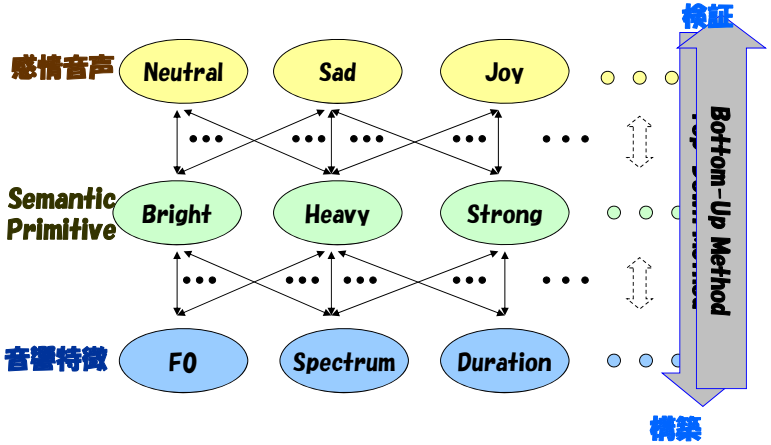
基本コンセプト：怒った声はどんな声？

- 高域パワーが○○dB大きくなった声
  → 正しい，しかし…．
- 大きな声，甲高い声 etc.
  → こちらが多い？
- 感情音声では



聴覚印象の多層モデル

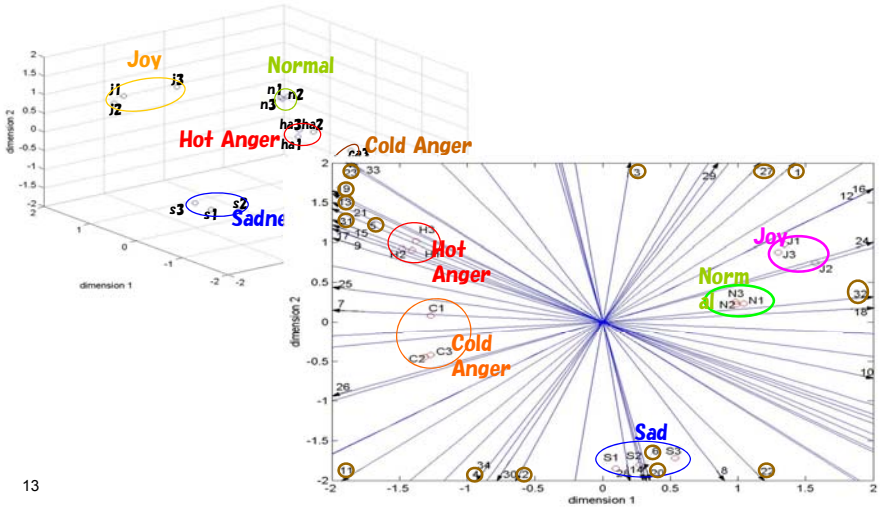
- 非言語情報　⇔　物理量を扱う信号処理
  → 表現語を介して結びつけるのが自然



モデルの構築： 感情音声

- 5つの感情:
  - Natural, Joy, Sad, Cold Anger, and Hot Anger
    プロの声優（女声）による演じられた感情音声．富士通研究所提供．
- 実験1
  - 感情音声データベース中の音声から感情は知覚されるか？
- 実験2
  - 多次元尺度構成法（ＭＤＳ）により，感情知覚空間を構成
- 実験3
  - モデルを構築するために表現語の候補を選び出す

実験２　感情知覚空間の構築
実験３　Semantic Primitive の選択



17個のSemantic Primitiveを選択

Semantic Primitives	
bright	monotonous
dark	heavy
high	clear
low	noisy
strong	quiet
weak	sharp
calm	fast
unstable	slow
well-modulated	

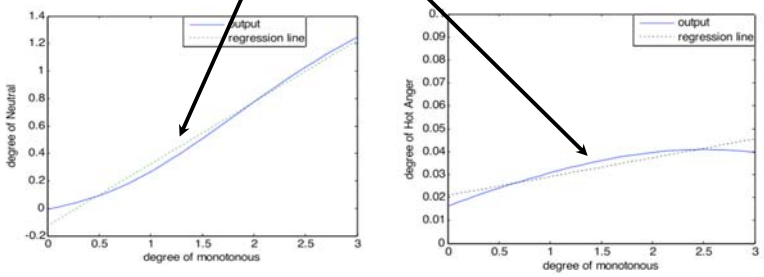
Fuzzy Interface System (FIS)の入出力関係

入力：17個のSemantic primitiveの強さ，**出力**：1つの感情の強さ

回帰直線の傾きを議論

傾きが正ならば，正の相関あり

傾きの大きさ（絶対値）が大きければ，関係が深い



FISの入出力関係：回帰直線の傾き

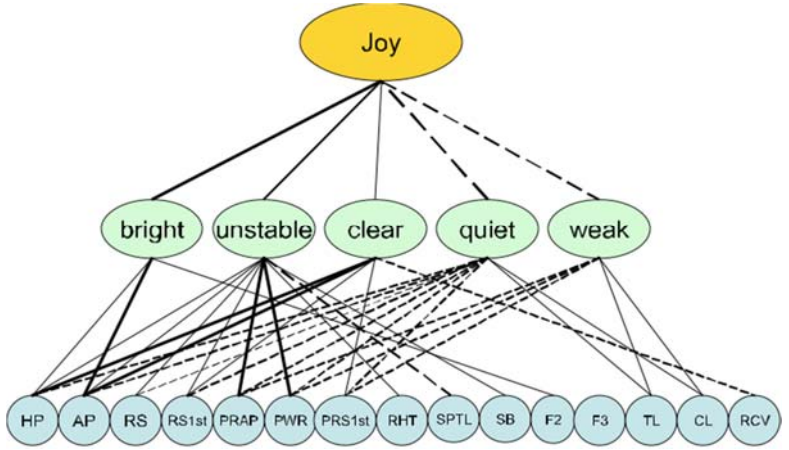
Neutral		Joy		Cold Anger		Sadness		Hot Anger	
PF	S	PF	S	PF	S	PF	S	PF	S
heavy	-0.329	quiet	-0.039	slow	-0.231	sharp	-0.079	calm	-0.063
weak	-0.181	weak	-0.036	monotonous	-0.073	strong	-0.049	quiet	-0.047
clear	0.127	unstable	0.063	fast	0.153	weak	0.065	unstable	0.120
monotonous	0.270	bright	0.101	heavy	0.197	heavy	0.074	well-modulated	0.124
calm	0.103	clear	0.034	well-modulated	0.091	quiet	0.057	sharp	0.103

音響特徴の選択

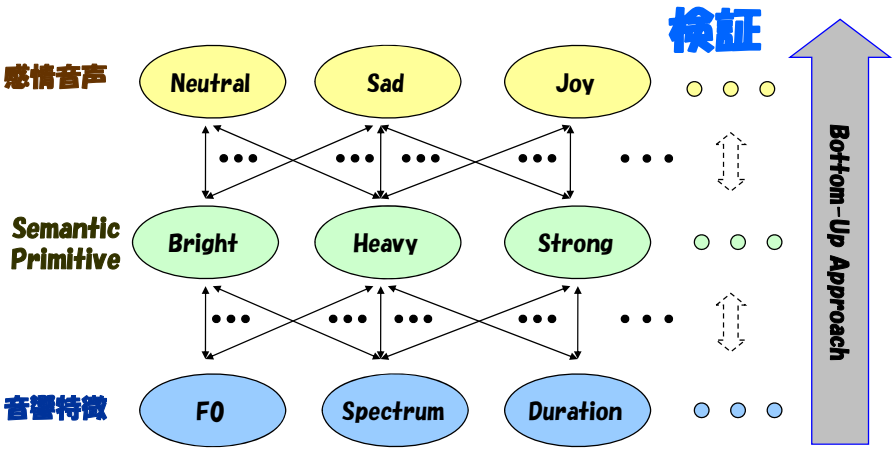
- 27個の音響特徴を計測
  - F0 contour : 8 features, Power envelope : 8 features, Spectrum : 5 features, Time duration : 6 features
- 相関分析
- 相関値0.6以上を採用
- 表現語と相関の高い16個の音響特徴を選択

	PF	bright	dark	high	low	strong	weak	calm	unstable	well mo-	mono-tonous	heavy	clear	noisy	quiet	sharp	fast	slow
F0	RS	0.44	-0.64	0.70	-0.60	0.56	-0.54	-0.74	0.67	0.54	-0.32	-0.40	0.44	0.63	-0.67	0.59	0.50	-0.56
	HP	0.69	-0.88	0.90	-0.89	0.42	-0.56	-0.72	0.67	0.50	-0.18	-0.73	0.74	0.60	-0.73	0.44	0.42	-0.62
	AP	0.71	-0.88	0.87	-0.91	0.33	-0.54	-0.66	0.60	0.41	-0.10	-0.78	0.76	0.52	-0.70	0.34	0.35	-0.62
power envelope	RS1st	0.50	-0.79	0.77	-0.78	0.45	-0.58	-0.67	0.60	0.42	-0.10	-0.61	0.66	0.57	-0.72	0.47	0.24	-0.51
	PRAP	0.31	-0.67	0.62	-0.56	0.73	-0.67	-0.77	0.73	0.55	-0.26	-0.30	0.47	0.76	-0.78	0.73	0.31	-0.55
	PWR	0.43	-0.74	0.74	-0.65	0.70	-0.66	-0.80	0.78	0.59	-0.27	-0.41	0.57	0.76	-0.79	0.69	0.38	-0.57
	PRS1st	0.48	-0.80	0.64	-0.70	0.45	-0.78	-0.61	0.51	0.27	-0.01	-0.56	0.64	0.44	-0.78	0.42	0.37	-0.64
	RHT	-0.10	-0.05	0.29	0.00	0.68	-0.14	-0.55	0.67	0.52	-0.41	0.24	-0.10	0.72	-0.29	0.68	0.36	-0.16
spectrum	F1	0.41	-0.64	0.59	-0.60	0.25	-0.39	-0.49	0.52	0.17	0.10	-0.49	0.47	0.43	-0.52	0.29	0.30	-0.29
	F2	0.60	-0.41	0.50	-0.56	-0.31	0.07	-0.11	0.07	0.08	0.05	-0.66	0.44	-0.03	-0.09	-0.27	0.11	-0.06
	F3	0.60	-0.47	0.61	-0.54	0.01	-0.15	-0.33	0.33	0.33	-0.16	-0.55	0.49	0.23	-0.29	0.02	0.27	-0.10
	SPTL	-0.29	0.49	-0.65	0.53	-0.48	0.17	0.62	-0.71	-0.49	0.21	0.30	-0.32	-0.72	0.42	-0.53	-0.23	0.24
	SB	0.27	-0.44	0.63	-0.48	0.49	-0.16	-0.55	0.66	0.55	-0.29	-0.28	0.28	0.68	-0.39	0.51	0.20	-0.31
time duration	TL	-0.26	0.42	-0.25	0.30	-0.41	0.69	0.52	-0.28	-0.19	0.19	0.21	-0.28	-0.22	0.63	-0.39	-0.59	0.80
	CL	-0.36	0.64	-0.39	0.53	-0.34	0.71	0.50	-0.32	-0.10	-0.04	0.47	-0.44	-0.29	0.71	-0.31	-0.37	0.59
	RCV	-0.41	0.78	-0.47	0.71	-0.14	0.58	0.29	-0.23	0.02	-0.32	0.66	-0.66	-0.27	0.58	-0.12	0.00	0.28

知覚モデル - Joy



モデルの検証



音響特徴と基礎的印象との関係

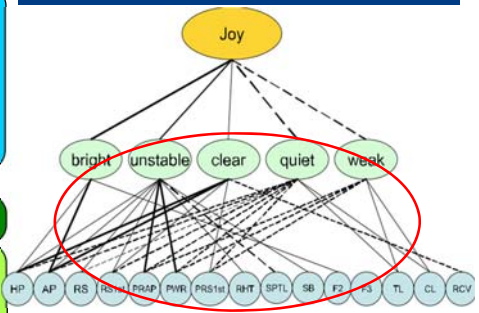
1. Base Rule

目的
　音響特徴の組み合わせの検証：選ばれた特徴と重みは適切か？
構成
　分析結果をそのまま使用

2. Intensity Rule

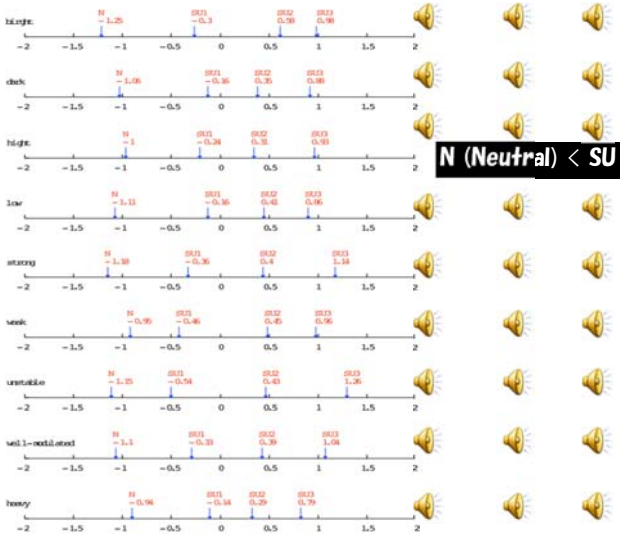
目的
　音響特徴とSemantic primitiveとの関係の検証：音響特徴がどのように基礎的印象に影響を与えるのか？
構成
　base rule を制御

Will bright voice really be affected by HP, AP, and F2?



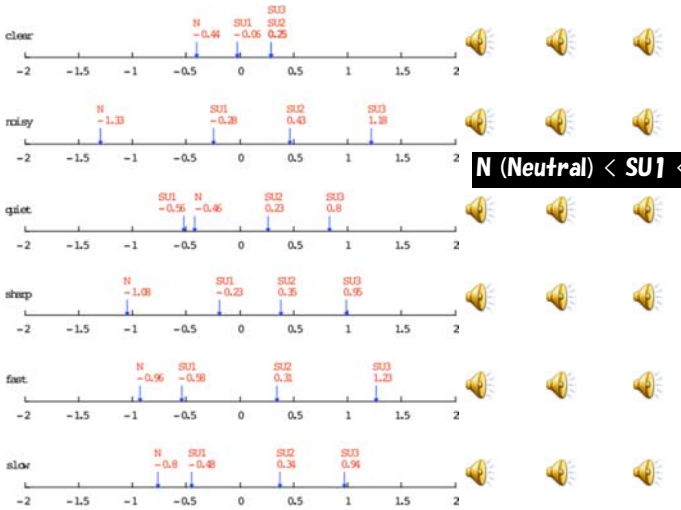
Is a voice sound brighter?

音響特徴と基礎的印象との関係デモ（1）



N (Neutral) < SU1 < SU2 < SU3

音響特徴と基礎的印象との関係デモ（2）



N (Neutral) < SU1 < SU2 < SU3

基礎的印象と感情の関係

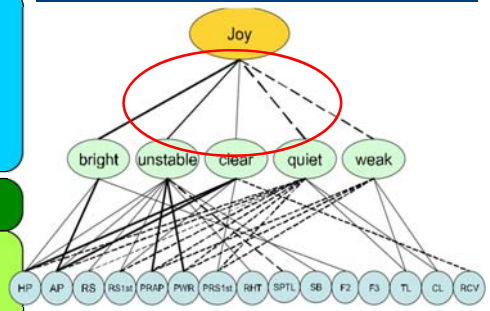
1. Base Rule

目的
Semantic primitiveの組み合わせの検証：選択と重みは適切か？
構成
FISでの分析結果をそのまま使用

2. Intensity Rule

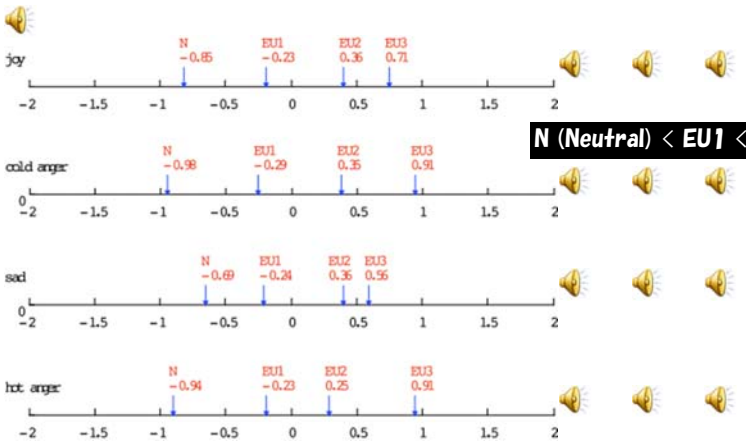
目的
Semantic primitiveと感情の関係の検証： Semantic primitiveがどのように感情の知覚に影響を与えるか？
構成
base rule を制御

Will Joy voice really sound bright, unstable, and clear, but not quiet and weak?



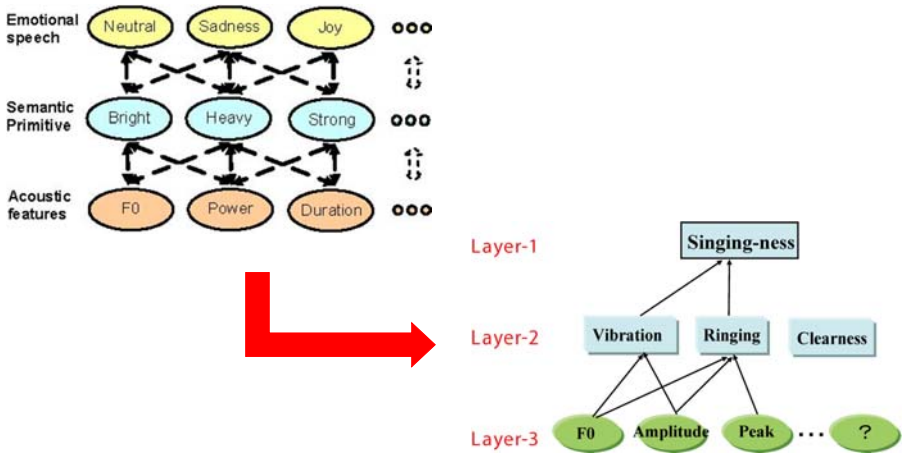
Is a brighter voice sound more Joyful ?

基礎的印象と感情の関係：デモ



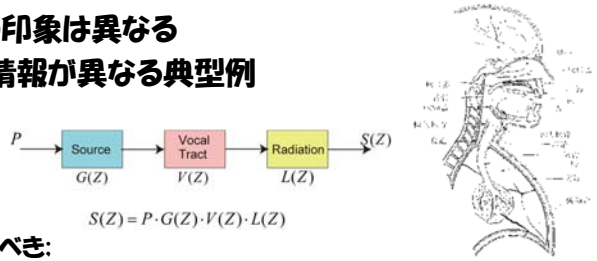
N (Neutral) < EU1 < EU2 < EU3

多層構造知覚モデルの歌声への応用



歌声の合成

- 歌声：歌詞（言語情報）は朗読しても歌っても同じ
- しかし，聞く側の印象は異なる
  - → 非言語情報が異なる典型例



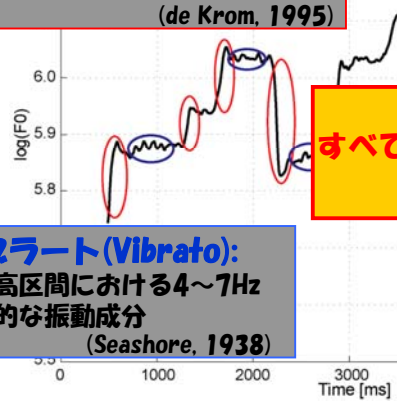
$$S(Z) = P \cdot G(Z) \cdot V(Z) \cdot L(Z)$$

- 例：歌声の合成
  - 次の点を考慮すべき：
    1. 声帯振動（メロディ：F0 Control, U/V ratio）
    2. 声道の構え（歌詞，響き：Spectra controll），および
    3. リズム（Duration control）

声帯振動：歌声に含まれる動的変動成分

オーバーシュート(Overshoot):
ポルタメント：傾斜を持った音高変化
エクステント：音高変化直後に目的音高を越える振動成分
(de Krom, 1995)

プレパレーション(Preparation):
音高が変化する直前に変化とは逆方向に瞬時に振れる成分

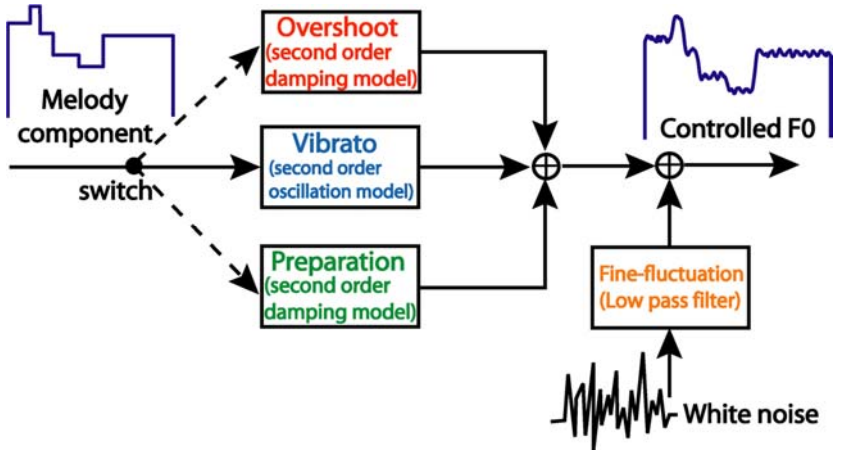


すべての成分が自然さにかかわる
(Saitou, Unoki and Akagi, ICAD2002)

ヴィブラート(Vibrato):
同一音高区間における4～7Hzの周期的な振動成分
(Seashore, 1938)

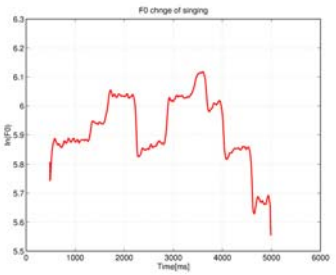
微細変動(Fine fluctuation):
発声区間全体に観測される不規則で細かい振動成分
(Akagi *et al*,2000)

F0制御モデル：概要

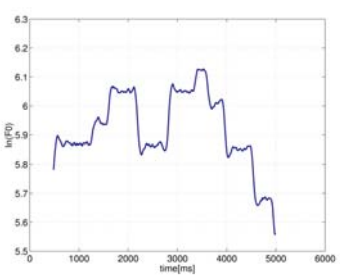


F0制御モデル：デモ

- 抽出F0 vs. 合成F0
  - フィルタは，実歌声からSTRAIGHTで抽出



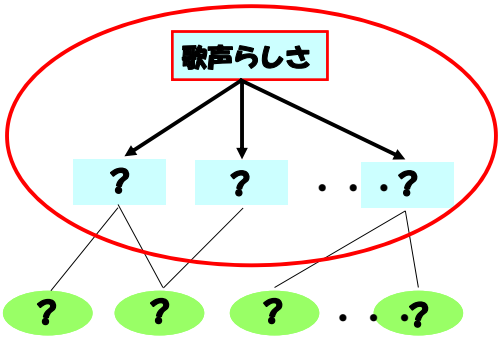
実F0包絡



モデルによる合成F0包絡

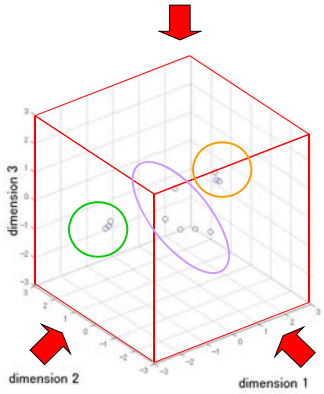
声道の構え：「歌声らしさ」のための三層モデル

- 第1層から第2層へ
  - どのような心理的要因が「歌声らしさ」に関係しているのか？
  - 多次元尺度構成法（Multi-Dimensional Scaling (MDS)）および多重回帰分析法（Multiple Regression Analysis (MRA)）を使用

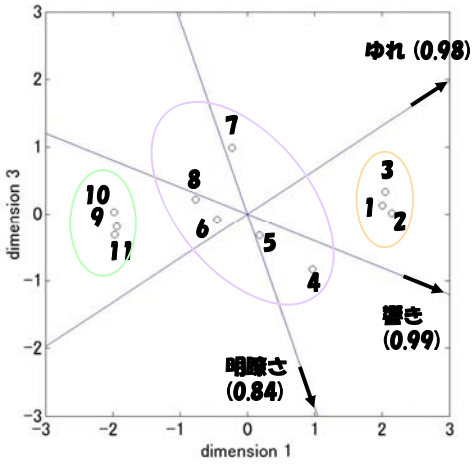


MDSとMRAの結果

group1(1,2,3)
group2(4,5,6,7,8)
group3(9,10,11)

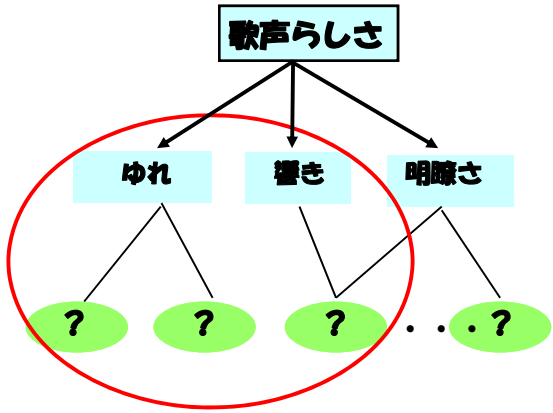


MDS: stress 5.4%

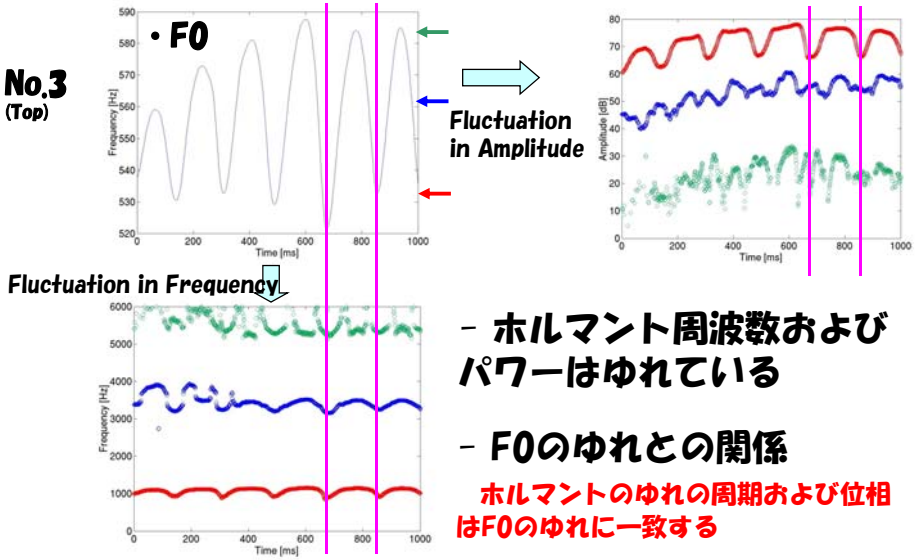


第2層と第3層の関係

第1層
第2層
第3層



ホルマントの変動



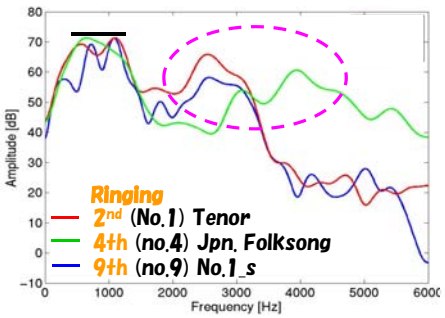
- ホルマント周波数およびパワーはゆれている
- F0のゆれとの関係

ホルマントのゆれの周期および位相はF0のゆれに一致する

響きに関係する音響特徴

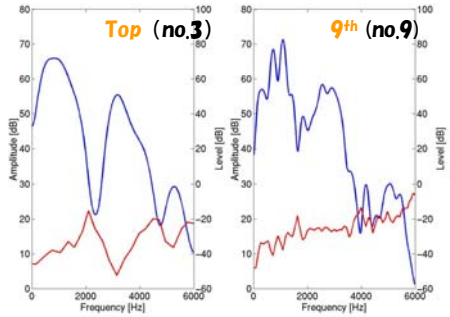
- 3 kHzあたりに存在する特徴的なピーク（*singers' formant*）, Sundberg, 1978

- 長時間スペクトル



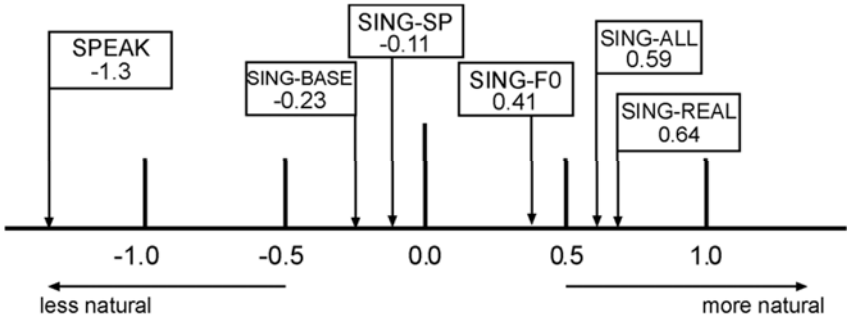
3 kHzあたりのピーク

- 非周期成分の比（Aperiodicity index）



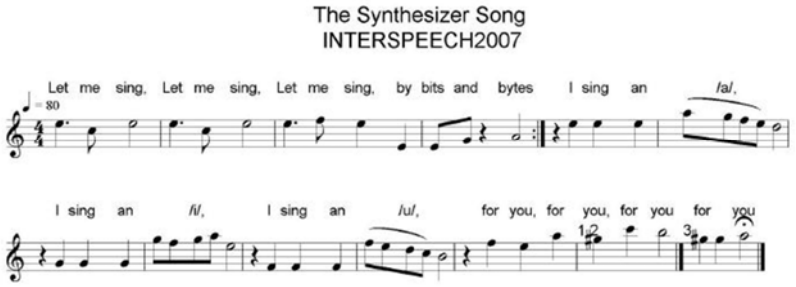
ピークで周期性が高い

話声から歌声へ：デモ(#1)



	SPEAK	SING-BASE	SING-SP	SING-F0	SING-ALL
Japanese children's song (female)	🔊	🔊	🔊	🔊	🔊
Classical singing (male)	🔊	🔊	🔊	🔊	🔊

デモンストレーション(#2)



♪ Speaking voice (input): 🔊(male) 🔊(female)

♪ Synthesized singing voice: 🔊(male) 🔊(female) 🔊(chorus)

We took the first place in SINGING SYNTHESIS CHALENGE held in the InterSpeech2007.

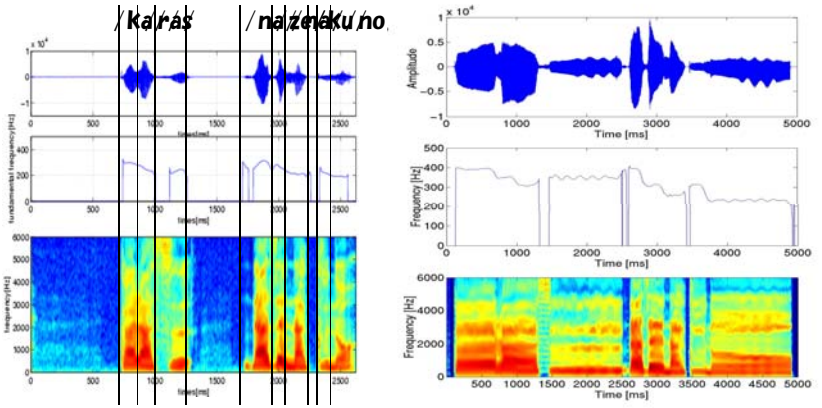
デモンストレーション(#3)

- 他の声区

	Speaking	Singing
Speaker-A		
Speaker-B		

デモンストレーション(#4)

- 話声　　合成歌声



歌声

ここまでのまとめ：表現豊かな音声の合成

- 音声の声質に関する知覚モデルの提案
- モデルの応用：２つの例
  1. 感情音声の合成
  2. 歌声の合成

- 発展：
  1. 感情音声の合成：
     声優？ (Computer graphics --> Computer audition)
  2. 歌声の合成：
     打倒・初音ミク!:

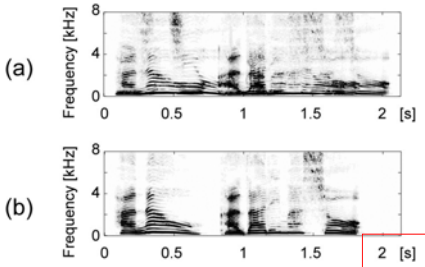
カクテルパーティ効果のモデル化

聞き耳をたてるコンピュータ

カクテルパーティ効果のモデル化

- **カクテルパーティ効果**
  混在する音の中から聴きたい音を聞分ける能力

デモンストレーション



キー情報：(b)の音声

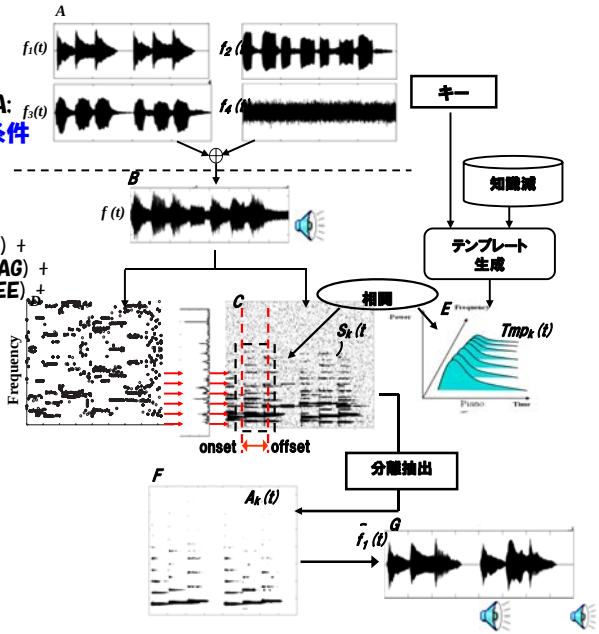
聞き耳をたてる
狙った音だけを聞き取る

➡ モデル化

モデルのデモ

Auditory Scene Analysis (ASA: Bregman 1993)の4つの制約条件をモデル化

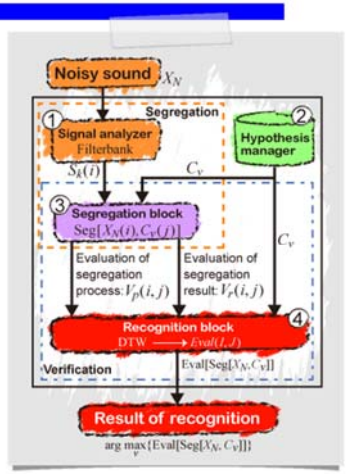
- 混合音:
  - ピアノ(チューリップCDECDE) +
  - フルート(キラキラ星CCGGAAG) +
  - ヴァイオリン(チョウチョGEEGEE) +
  - 白色雑音(背景雑音)
- キー: Piano＋
  楽譜情報(CDECDE)
- 条件: SNR=0 [dB]



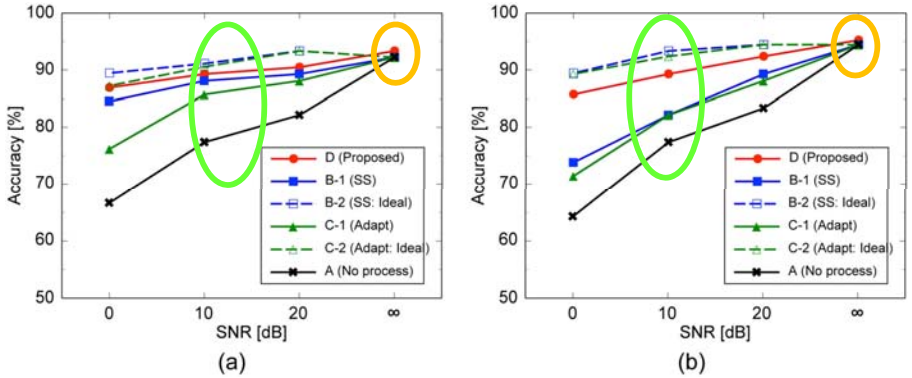
機械による音声認識への応用

- モデルのコンセプトを音声認識システムへ応用
  1. 信号処理ブロック
  2. 仮説生成マネージャー
  3. 分凝ブロック
  4. 認識ブロック

＊4つの心理的な規則を適用



実験結果



Babble noise　　Machinegun

プライバシー保護

カクテルパーティ効果を逆手にとって

- 原理
  - 分凝と融合：
    - 分離できず，一つに聞こえてしまう音とは？
- 銀行の窓口，病院の診察，等でのプライバシー保護
  - グローリー(株)と共同研究
  - セキュリティショー(2007年3月)
  - 日経トレンディ(2009年)
  - 発売(2011年9月)
- 波及効果
  - 世界へ
    - 赤木正人他4名，音声処理方法と装置及びプログラム並びに音声システム，(共同研究：(株)グローリー)：特許取得(日，米，EU，中国，韓国)

話を聞かれても意味は悟らせない

“会話キャンセラー”で盗み聞きを防ぐ

●会話を聞き取れなくする仕組み

Ⓐ防聴音は、マイクが拾った会話からリアルタイムで生成される。会話がない状況では音は流れない。BGMなどで会話を覆い隠す方式とは決定的に違う

Ⓐついたて型の試作機。右の男性が電話する声は、この近距離でも左の女性には聞き取れない Ⓑ銀行などの窓口も想定

今後の改善点 それぞれの現場に合わせた微調整
実用化の時期 数カ月以内に試験運用を開始
ヒット予測 ★★

驚がないオープンスペースで、しかも周囲は人だらけ。でも、大事な話をしても聞かれる心配がない――。そんな不思議な場所が、年内にも一部の店などに出現するかもしれない。

資帳処理機を手がけるグローリーは、わずか1mの距離に人がいても、会話のプライバシーを“保護”できる技術を開発した。「聞こえなくする」のではない。同社の装置を設置すると、その場所での会話は、周囲の人にとって「聞こえるが、意味は不明」な状態になる。まるで知らない外国語を聞いているかのようだ。

秘密は、設置されたスピーカーから流れる「防聴音」。これと会話が混ざった音は人の声に似ているが、内容を全く聞き取れない。不快なノイズではないが、意味だけがわからなくなる。防聴音はマイクが拾った会話から自動生成。騒音を打ち消すノイズキャンセリングに近い仕組みという。会話する当人にとっては、自分たちの声よりも防聴音が小さいため、内容をしっかり聞き取れる。

大掛かりな工事は不要で、簡単な間仕切りなどを置くだけで効果が得られる。防音壁などの手段と比べて、導入の敷居が低い技術という。

銀行窓口や薬局のカウンターなど、立ち聞きが気になる場所は多い。技術はおおむね完成し、数カ月以内にオフィスなどで試験運用が始まる。年内の店舗への導入もありそうだ。

プライバシー保護

HOME 経済 記事 神戸新聞

経済

世界初、防聴音で会話保護 グローリー

グローリーが開発した会話保護システム。窓口での会話を集音して生成した「防聴音」を、間仕切りなどに設置した出力装置から出す＝姫路市下手野、グローリー

グローリー（姫路市）は、金融機関や薬局などの窓口で、会話の内容を第三者に聞き取れないようにする会話保護システムを開発、1日から発売した。言葉を認識しにくくする「防聴音」を会話に連動して発生させ、プライバシーを保護する技術は世界初という。

金融機関や医療関係の窓口では個人の資産や病歴などが語られるが、間仕切りつだけの場所が多い。防音壁は施工が大がかりになり、雑音をかぶせるサウンドマスキングは音量を上げる必要がある。

同社は生体認証などセキュリティー技術の開発を進めており、2007年から会話保護システムの製品化に着手。その音声固有の周波数を打ち消す「防聴音」を即時にかぶせ、言葉を認識させにくくすることに成功した。既に薬局などで試験導入されている。

システムは集音、音声処理、出力の各装置からなり、窓口と待合席の間などに出力装置を設置すれば、声らしきものは聞こえるが、内容は聞き取れないという。どの言語でも対応可能。定価60万円。レコード会社のビクターエンタテインメント（東京）のオーディオシステムを組み合わせ、リラックス効果が得られる音を出すタイプ（定価110万円）も売り出す。発売後1年で計1千台の販売を目標としている。（広岡磨璃）

(2011/09/02 07:36)

第二部まとめ：「聞き耳」

- 音情景理解についての知見およびそのモデルを紹介した
- 応用例
  - 楽器音抽出
  - 音声認識システム
  - 音声プライバシー保護法
- 音声プライバシー保護法
  - 会話音声と同時に音声の音韻性を曖昧にする防聴音を再生
  - 音情景理解がなるべく有効に働かない音環境を生成
  - 発話内容を不明瞭にする
  - 本手法を利用した製品がすでに発売され調剤薬局などに導入が始まっている