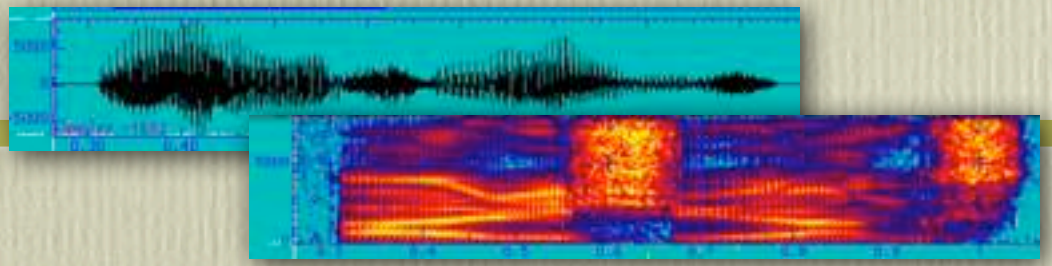
人文社会系研究科基礎文化研究専攻言語学専門分野

音響音声学

(Topics in Acoustic Phonetics)

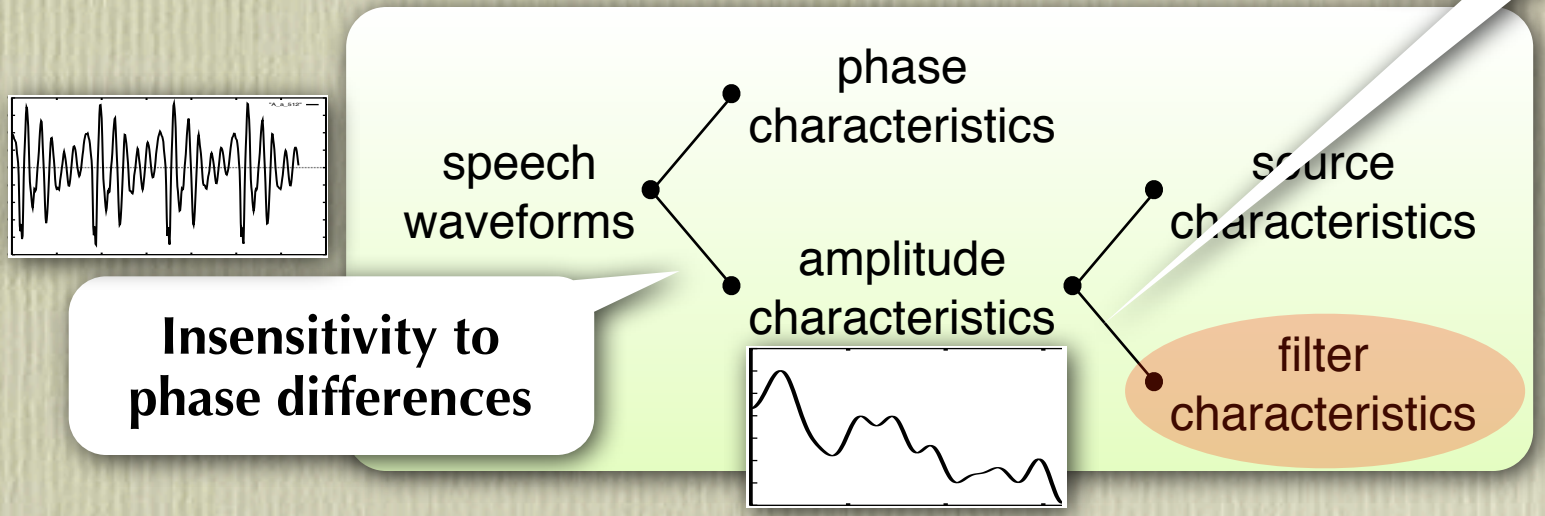


峯松 信明

工学系研究科電気系工学専攻

その情報を運ぶ媒体・音響特徴量

二段階の分離に基づく特徴量抽出



- スペクトル包絡(o)は何を運ぶのか？　　言・パラ言・非言

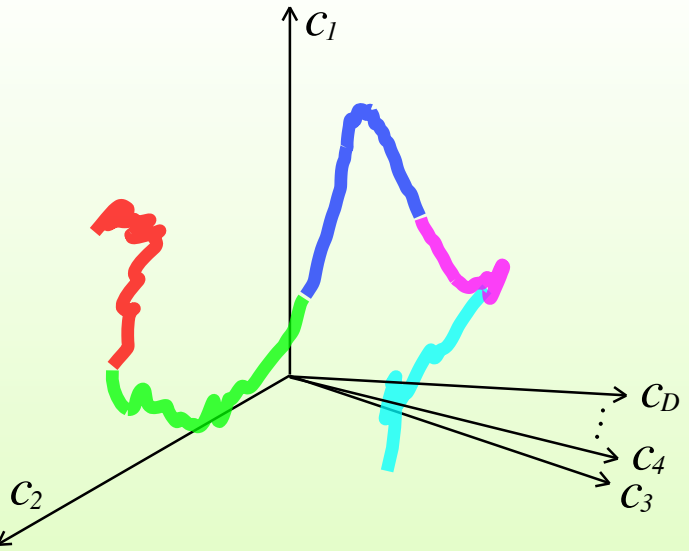
二つの音響モデル $P(o|w)$ と $P(o|s)$

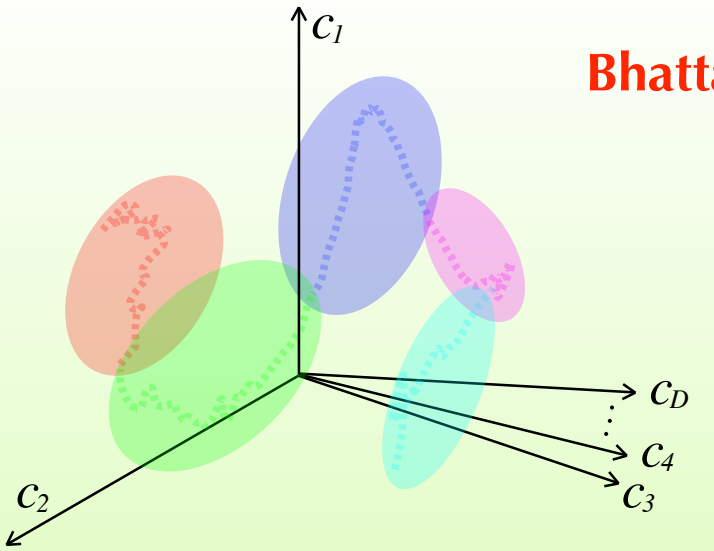
s = speaker
w = word

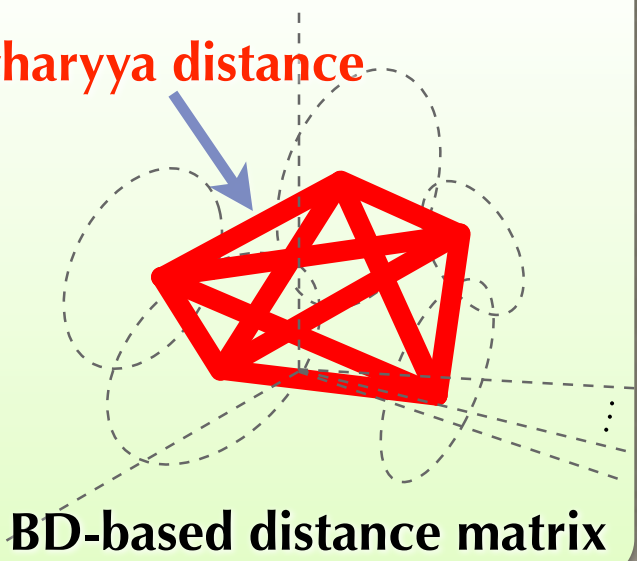
- 不特定話者の単語音響モデル
  - $P(o|w) = \sum_s P(o,s|w) = \sum_s P(o|w,s)P(s|w) \sim \sum_s P(o|w,s)P(s)$
- テキスト非依存の話者音響モデル
  - $P(o|s) = \sum_w P(o,w|s) = \sum_w P(o|w,s)P(w|s) \sim \sum_w P(o|w,s)P(w)$
  - 集めてしまえば「確率の定義」が見たくないものを隠してくれる。

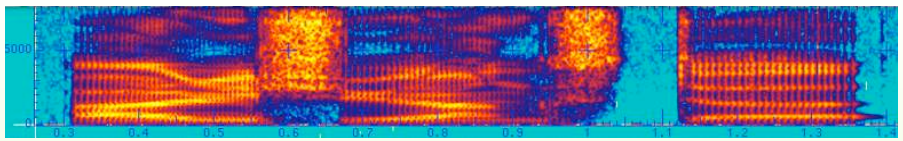
分布間距離群としての音声表象

ケプトラム系列 → 分布系列 → 距離行列









spectrogram (spectrum slice sequence)

cepstrum vector sequence

distribution sequence

Really speaker-independent features

Deep neural network [Hinton+'06, '12]

- Deeply stacked artificial neural networks
- Results in a huge number of weights
- Unsupervised pre-training and supervised fine-tuning

Findings in DNN-based ASR [Mohamed+'12]

- First several layers seem to work as extractor of invariant features or speaker-normalized features.
- Still difficult to interpret structure and weights of DNN physically.
  - Interpretable DNNs are becoming one of the hot topics [Sim'15].

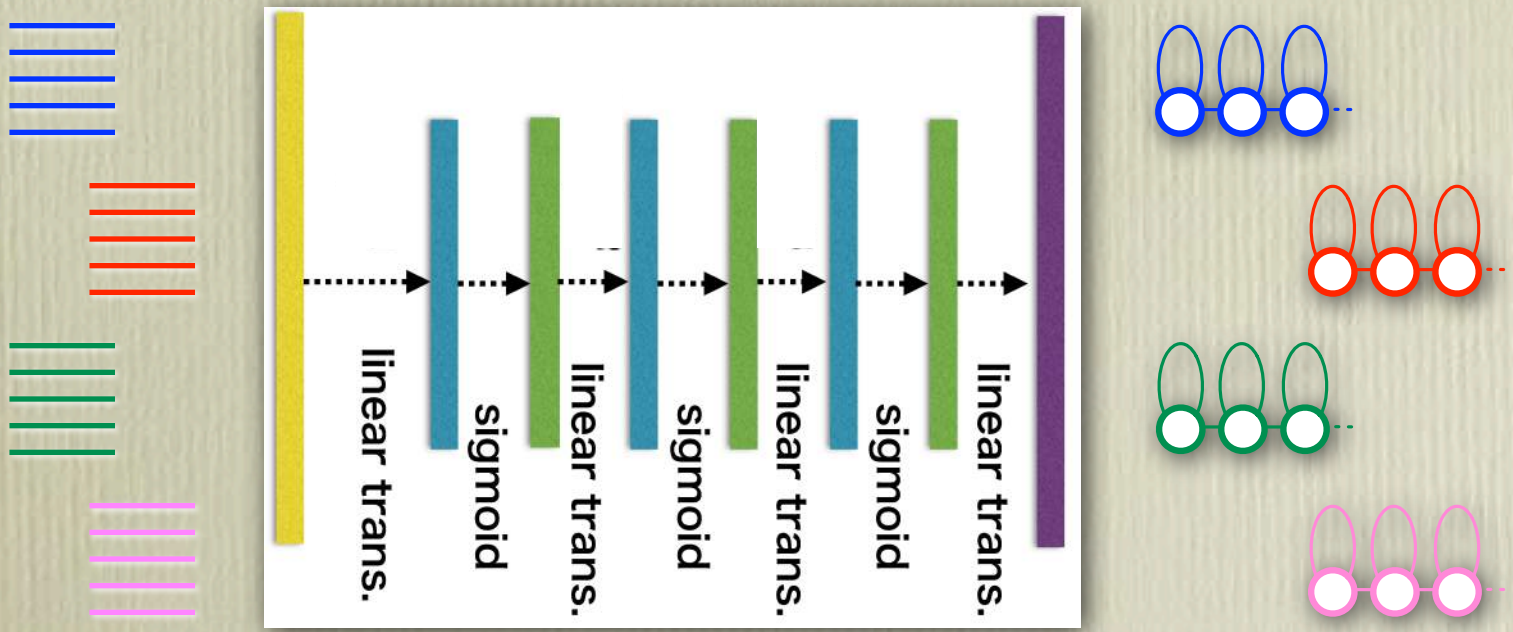
A simple question asked in tutorial talks of DNN

- "What are *really* speaker-independent features?"
  - Asked by N. Morgan at APSIPA2013 and ASRU2013

DNN as posterior estimator

General framework for training DNN

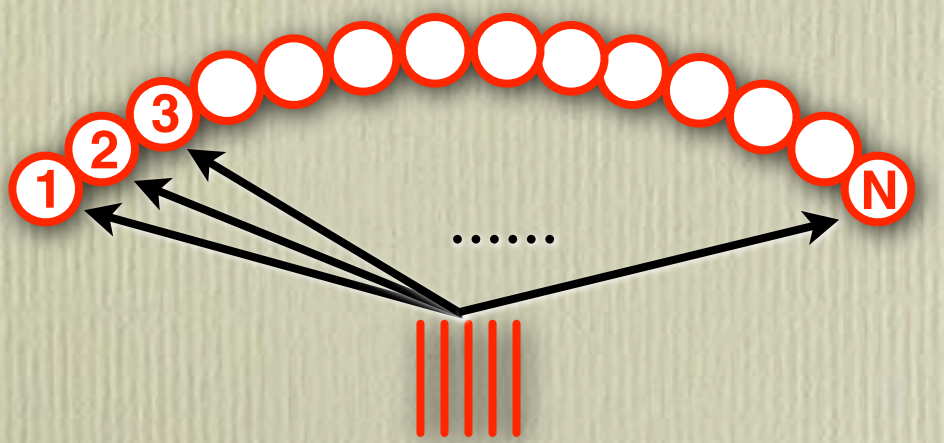
- Unsupervised pre-training and supervised training
  - In the latter training, speaker-adapted HMMs are used to prepare posteriors (=labels) for each frame of the training data.
- DNN is trained so that it can extract speaker-invariant features and estimate posteriors in a speaker-independent way.
- Output of DNN = posteriors (phoneme state posteriors in ASR)

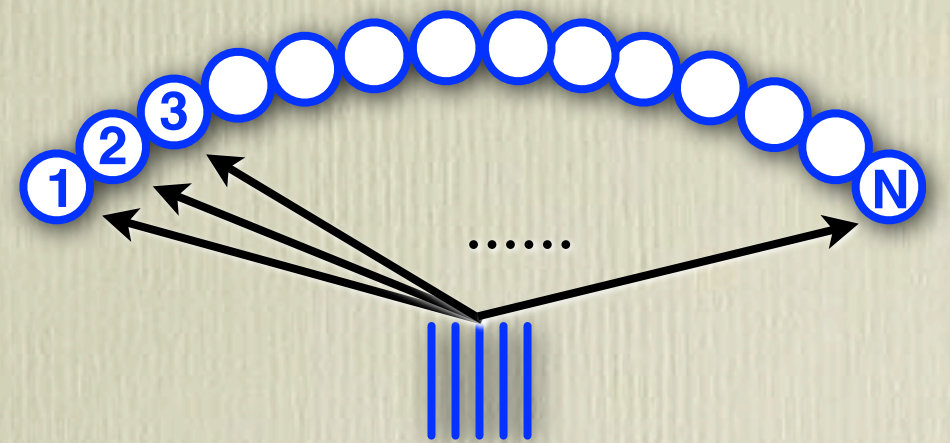


Posteriors = normalized similarities

Posteriors of $\{P(c_i|o)\}$

- $P(c_i|o) \propto P(o|c_i)P(c_i)$
- $\sum_i P(c_i|o) = 1.0$
- Can be interpreted as normalized similarity scores biased by priors.
- Output of DNN = normalized similarity scores to a definite set of speaker-adapted acoustic "anchors" of $\{c_i\}$.



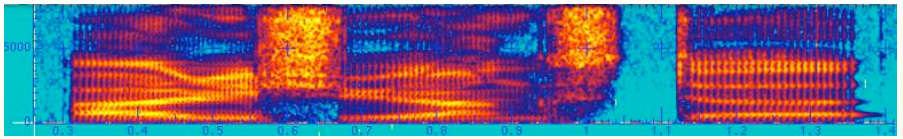


■ ■ : speaker-dependent ■ : speaker-independent(invariant)

- Similarities scores can be converted to **distances to "anchors"**.
  - Either of similarity matrix or distance matrix is used for clustering.

Distances to anchors

Speech structure extracted from an utterance



spectrogram (spectrum slice sequence)

cepstrum vector sequence

distribution sequence

Structure extraction for speakers ■ and ■

■ ■ : speaker-dependent

■ : speaker-independent(invariant)

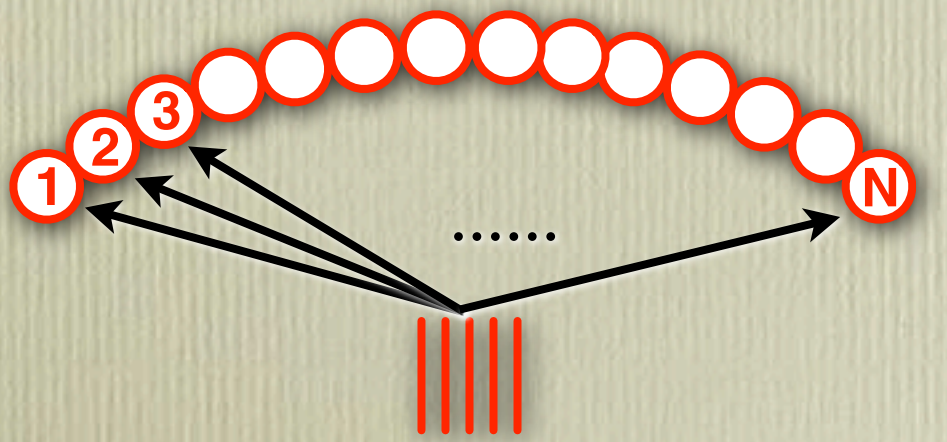
Invariant contrasts

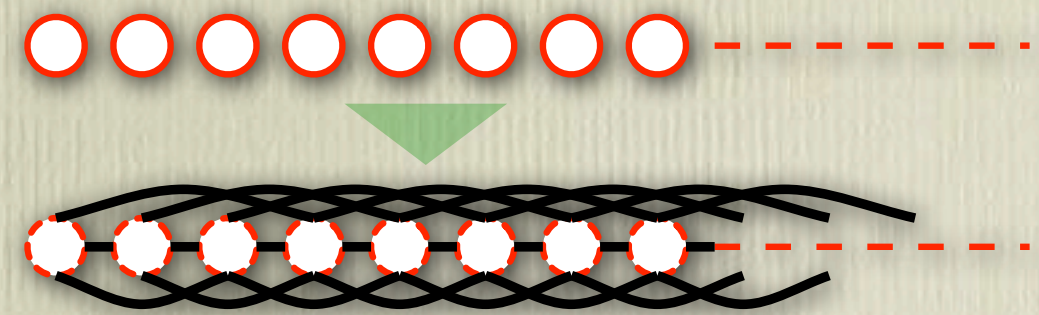
DNN as speaker-invariant contrast estimation

- Use of spk-dependent HMMs to prepare posterior labels
- A huge data to train DNN to guarantee spk-invariance

Str. extraction as speaker-invariant contrast detection

- Use of within-utterance acoustic events only
- Spk-invariance is guaranteed by invariant properties of f-div.





Origin and evolution of language

EVOLANG IX

KYOTO, JAPAN
13 - 16 March, 2012

Plenary Speakers:

- Noam Chomsky
- Minoru Asada
- Cedric Boeckx
- Terrence Deacon
- Simon Fisher
- Russell Gray

Abstract Submission

8 September, 2011

A MODULATION-DEMODULATION MODEL FOR SPEECH COMMUNICATION AND ITS EMERGENCE

NOBUAKI MINEMATSU

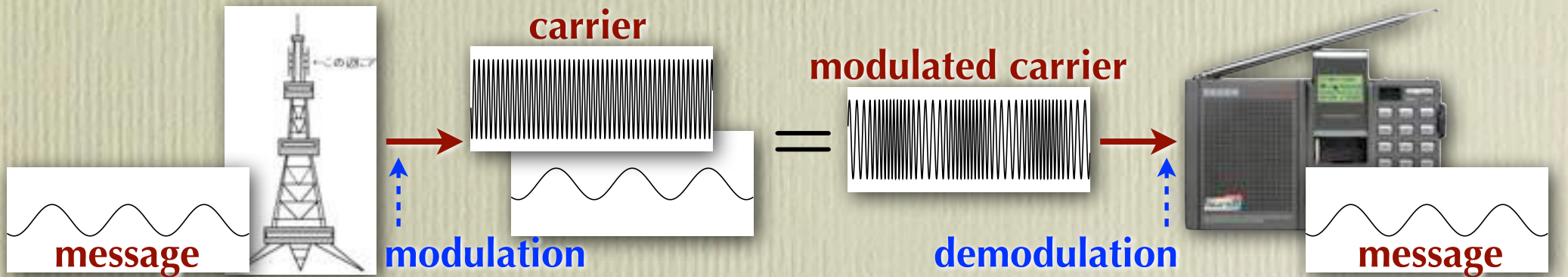
Graduate School of Info. Sci. and Tech., The University of Tokyo, Japan,
mine@gavo.t.u-tokyo.ac.jp

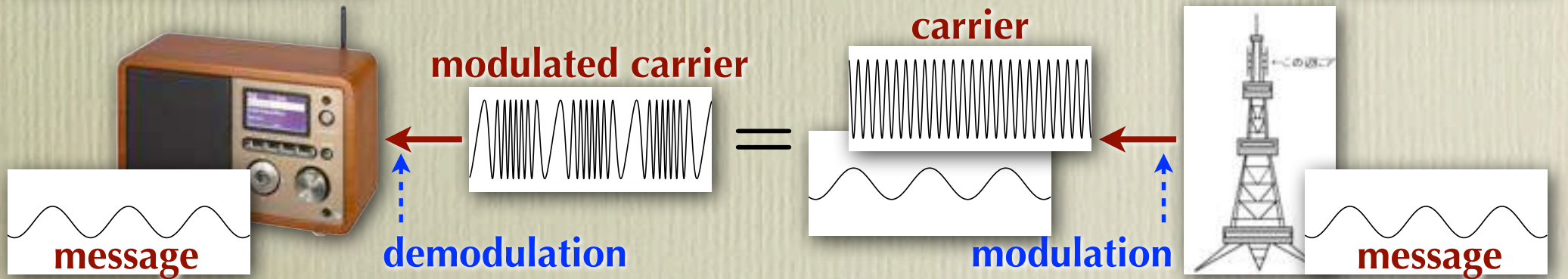
Perceptual invariance against large acoustic variability in speech has been a long-discussed question in speech science and engineering (Perkell & Klatt, 2002), and it is still an open question (Newman, 2008; Furui, 2009). Recently, we proposed a candidate answer based on mathematically-guaranteed relational invariance (Minematsu et al., 2010; Qiao & Minematsu, 2010). Here, transform-invariant features, f-divergences, are extracted from the speech dynamics in an utterance to form an invariant topological shape which characterizes and represents the linguistic message conveyed in that utterance. In this paper, this representation is interpreted from a viewpoint of telecommunications, linguistics, and evolutionary anthropology. Speech production is often regarded as a process of modulating the baseline timbre of a speaker's voice by manipulating the vocal organs, i.e., spectrum modulation. Then, extraction of the linguistic message from an utterance can be viewed as a process of spectrum *de*modulation. This modulation-demodulation model of speech communication has a strong link to known morphological and cognitive differences between humans and apes.

Modulation used in telecommunication

From Wikipedia

A musician modulates the tone from a musical instrument by varying its volume, timing and pitch. The three key parameters of a carrier sine wave are its amplitude ("volume"), its phase ("timing") and its frequency ("pitch"), all of which can be modified in accordance with a content signal to obtain the modulated carrier.

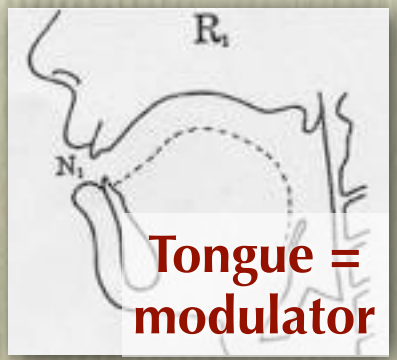




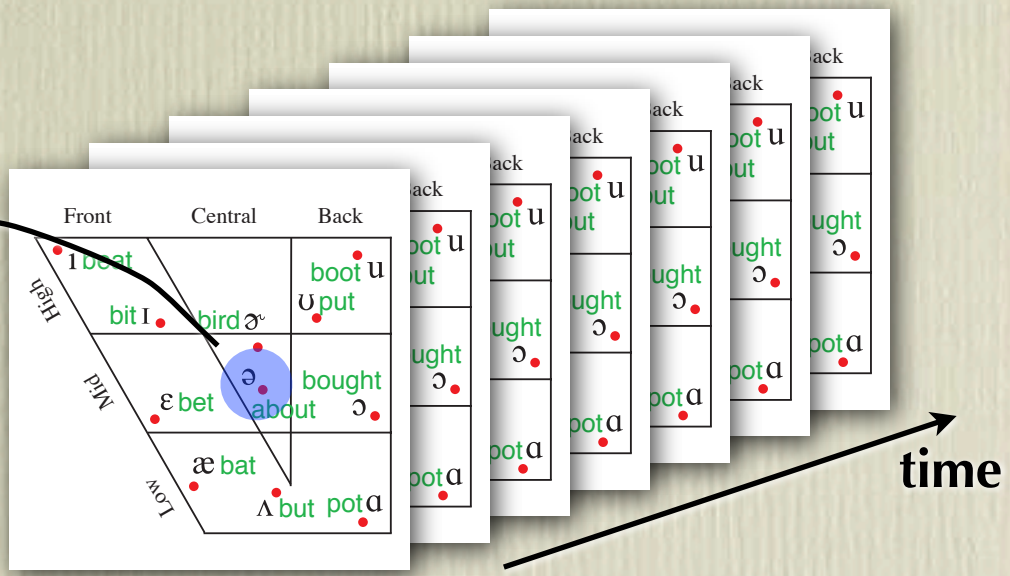
A way of characterizing speech production

Speech production as spectrum modulation

- Modulation in frequency (FM), amplitude (AM), and phase (PM)
  - = Modulation in pitch, volume, and timing (from Wikipedia)
  - = Pitch contour, intensity contour, and rhythm (= prosodic features)
- What about a fourth parameter, which is **spectrum (timbre)**?
  - = Modulation in spectrum (timbre) [Scott'07]
  - **= Another prosodic feature?**



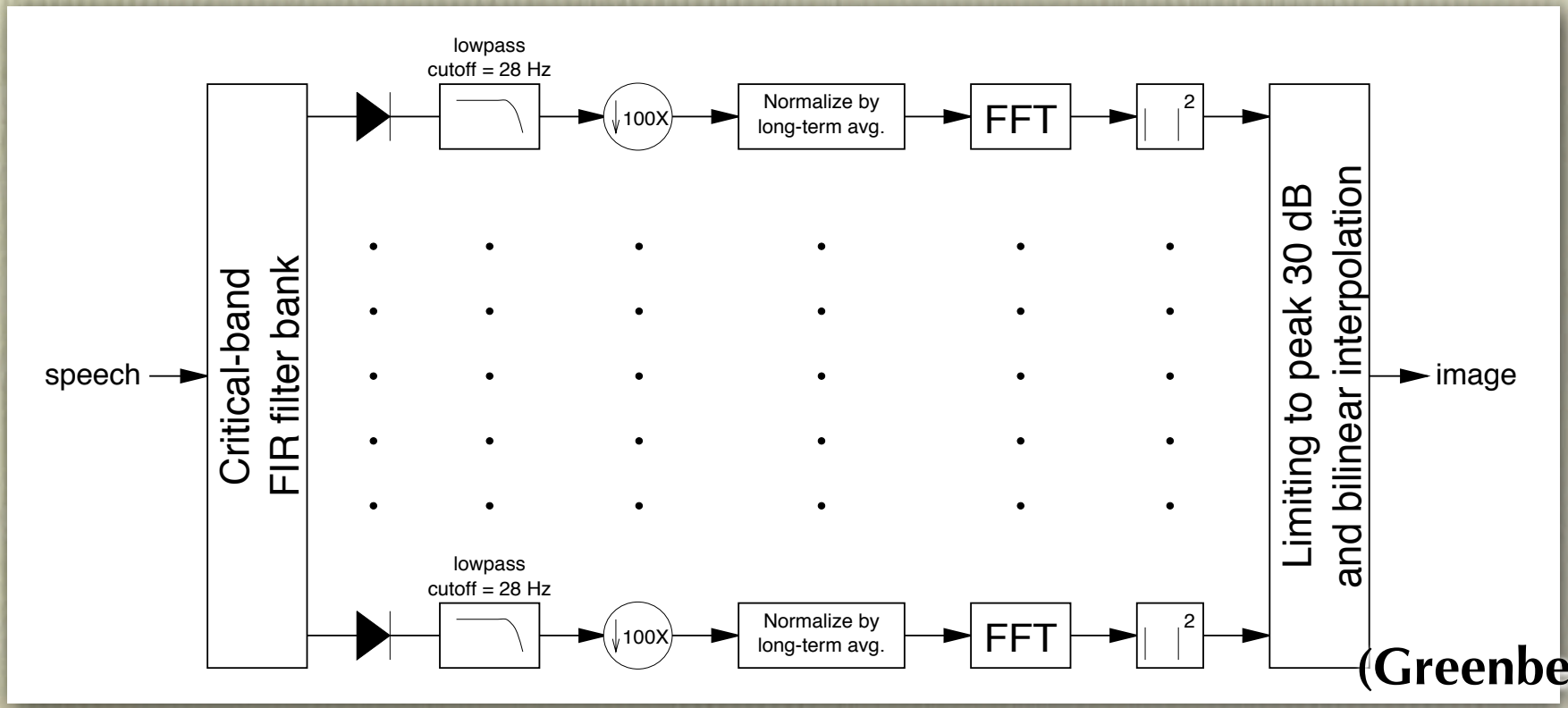
Schwa
= most lax
= most frequent
= home position
= spk.-specific baseline timbre



Modulation spectrum

Critical-band based temporal dynamics of speech

- "In pursuit of an invariant representation" (Greenberg'97)
- RASTA (=RelAtive SpecTrA, Hermansky'94)



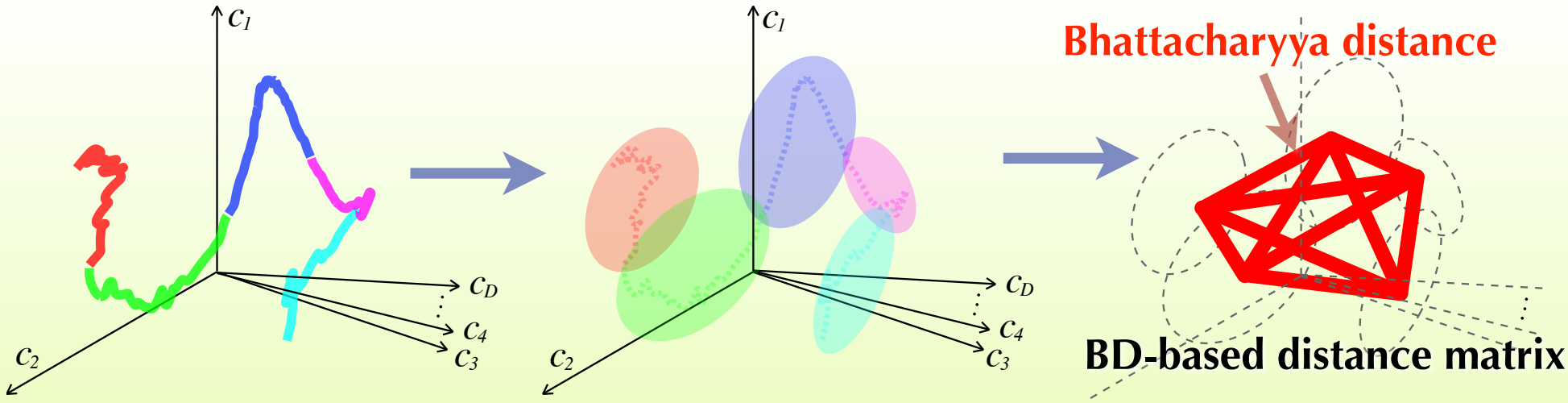
(Greenberg'97)

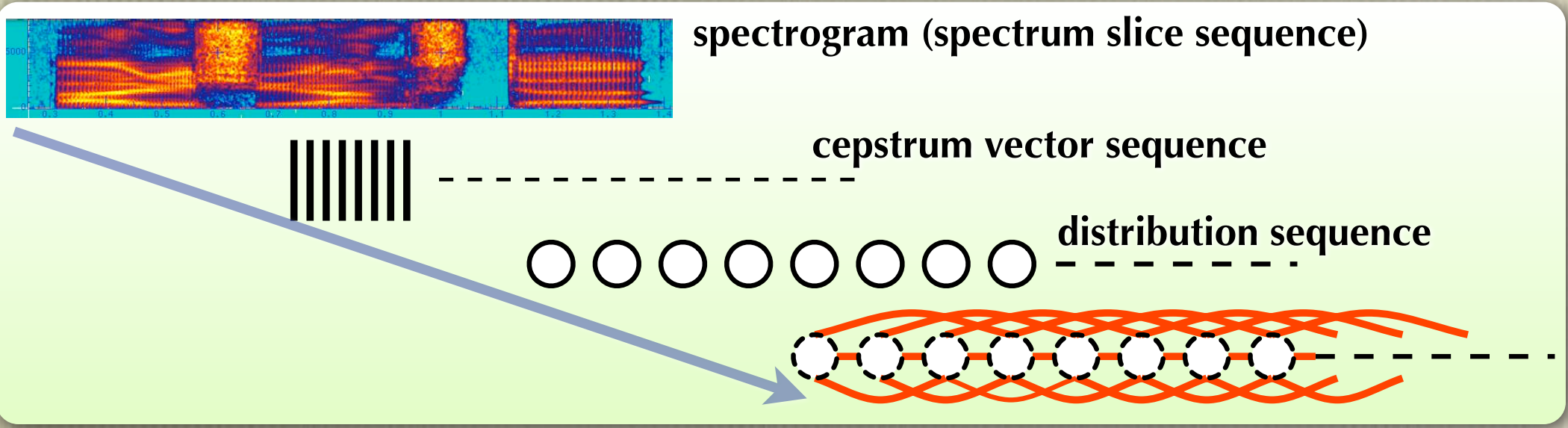
No mathematical proof for invariance

- Direction of a trajectory is rotated by VTL difference (Saito'08)

Invariant speech structure

Utterance to structure conversion using *f*-div. [Minematsu'06]



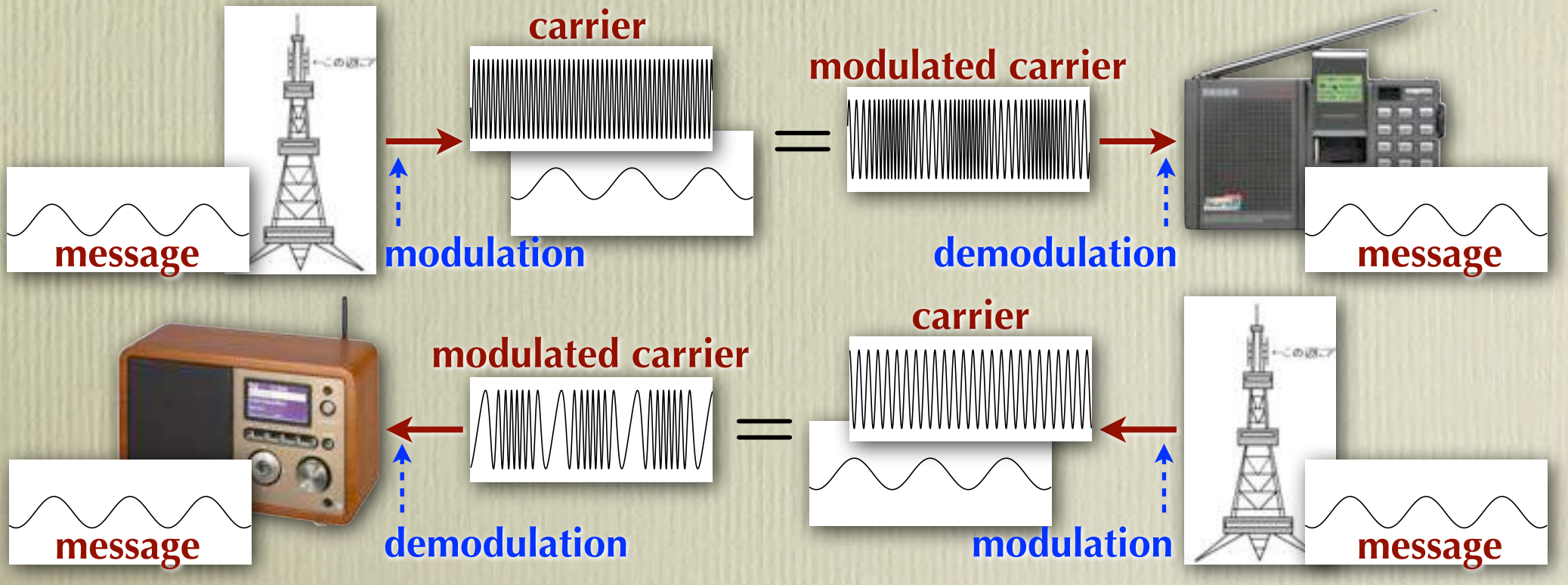


- An event (distribution) has to be much smaller than a phoneme.

Demodulation used in telecommunication

Demodulation in frequency, amplitude, and phase

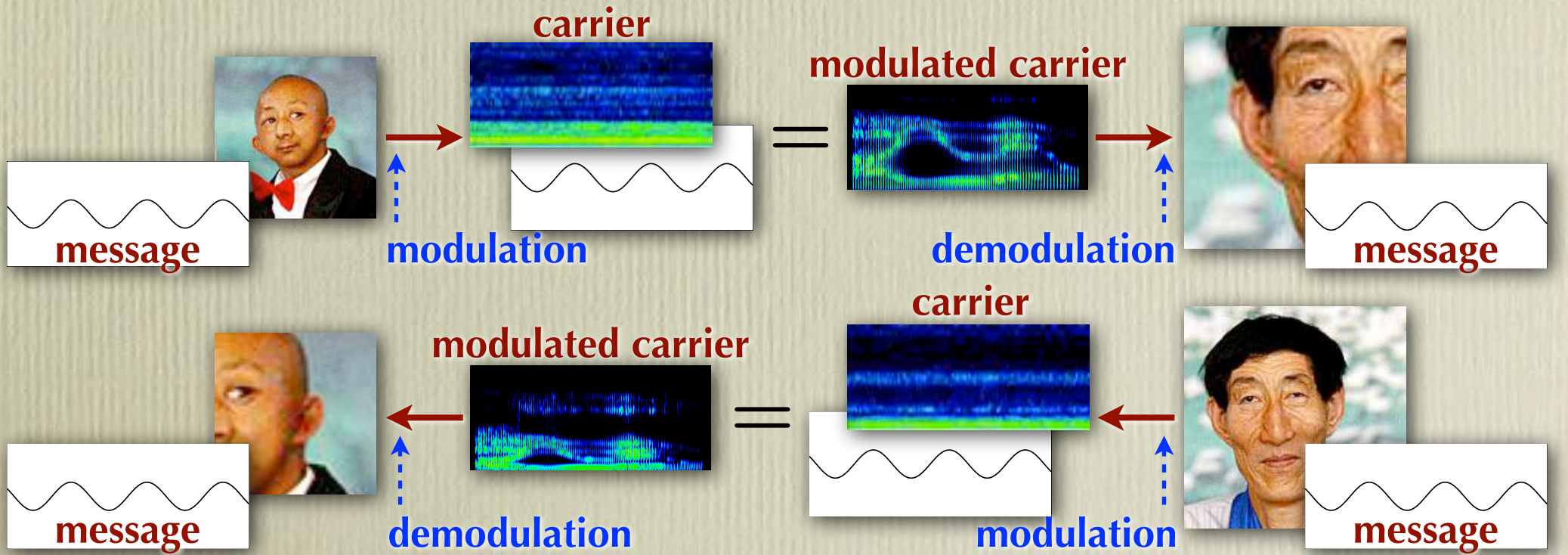
- Demodulation = a process of extracting a message intactly by removing the carrier component from the modulated carrier signal.
  - Not by extensive collection of samples of modulated carriers
  - (Not by hiding the carrier component by extensive collection)



Spectrum demodulation

Speech recognition = spectrum (timbre) demodulation

- Demodulation = a process of extracting a message intactly by removing the carrier component from the modulated carrier signal.
  - By removing speaker-specific baseline spectrum characteristics
  - Not by extensive collection of samples of modulated carriers
  - (Not by hiding the carrier component by extensive collection)



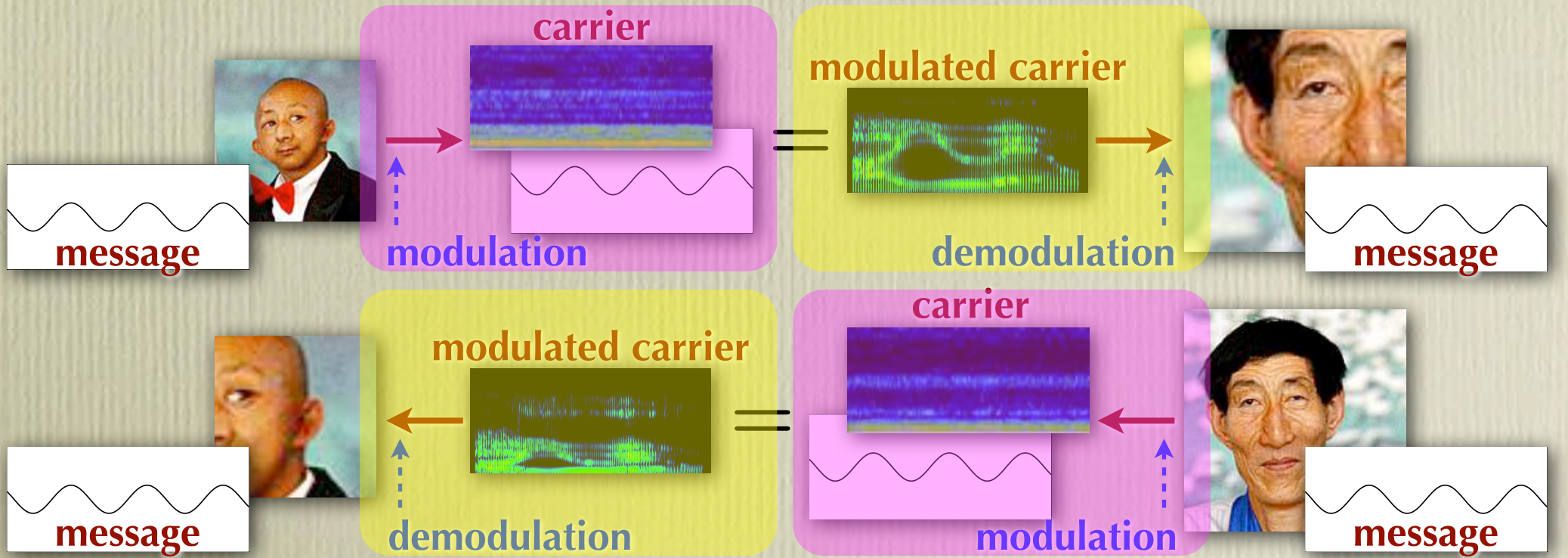
Two questions

Q1: Does the ape have a good modulator?

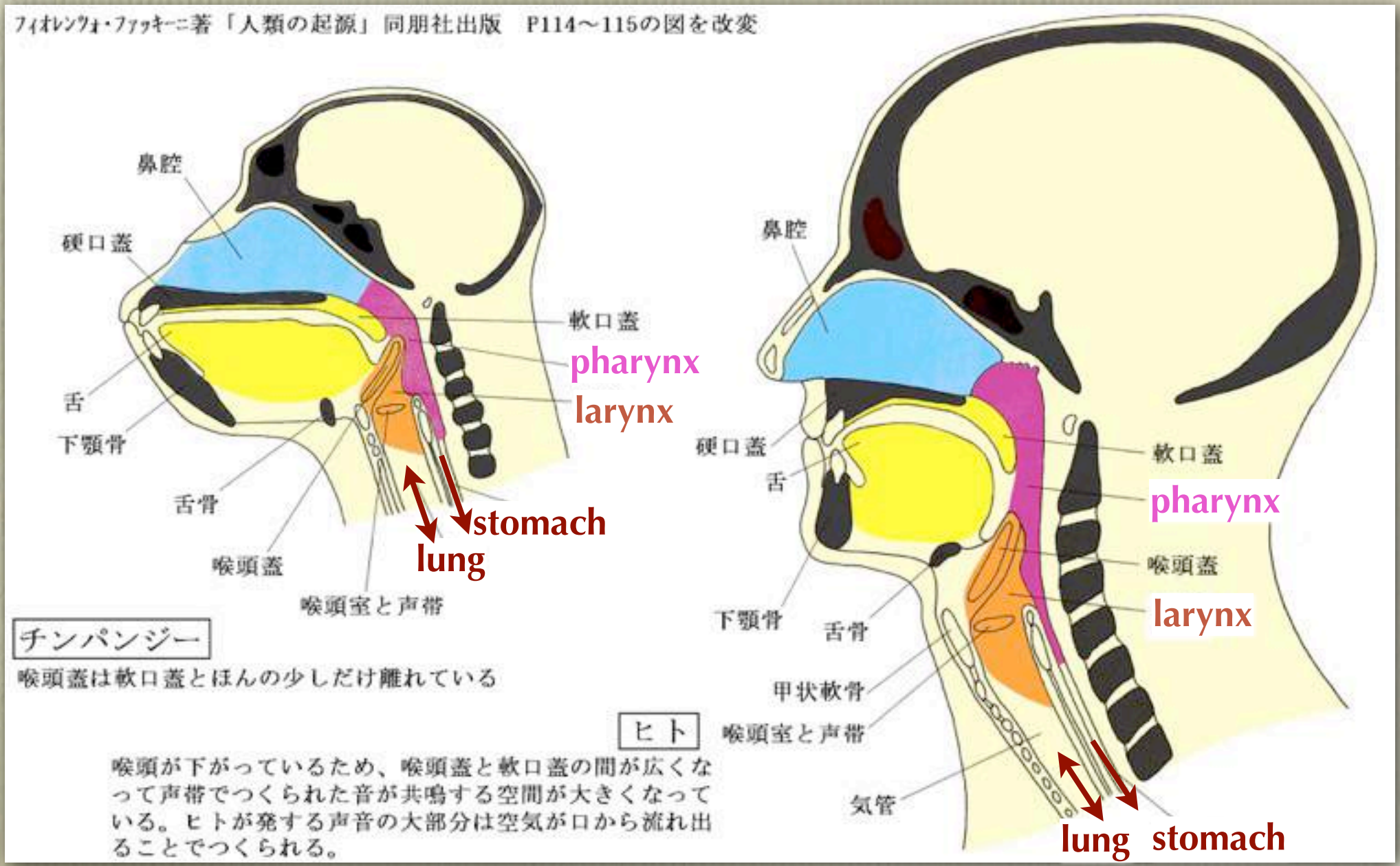
- Does the tongue of the ape work as a good modulator?

Q2: Does the ape have a good demodulator?

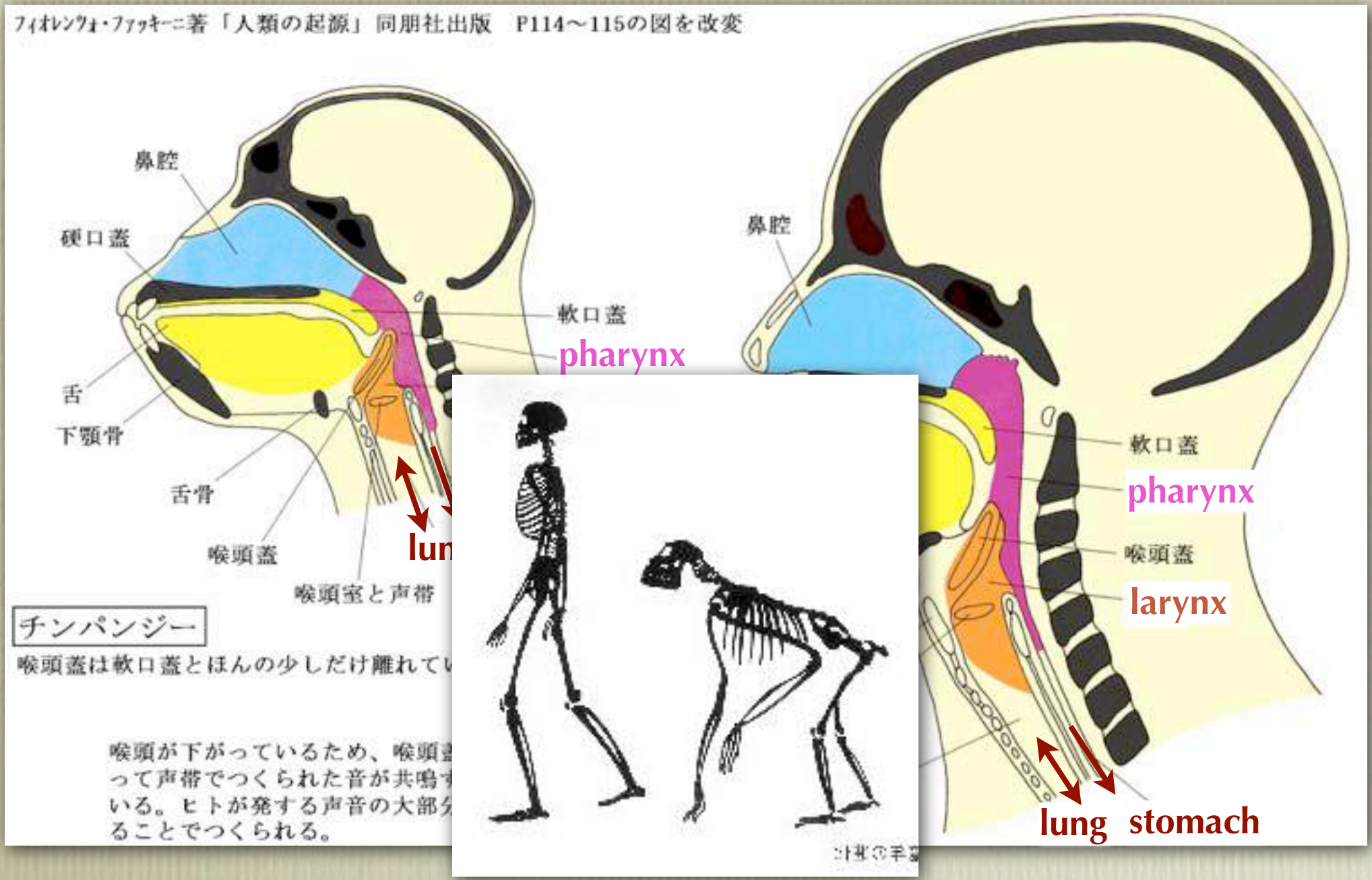
- Does the ear (brain) of the ape extract the message intactly?



Structural diff. in the mouth and the nose



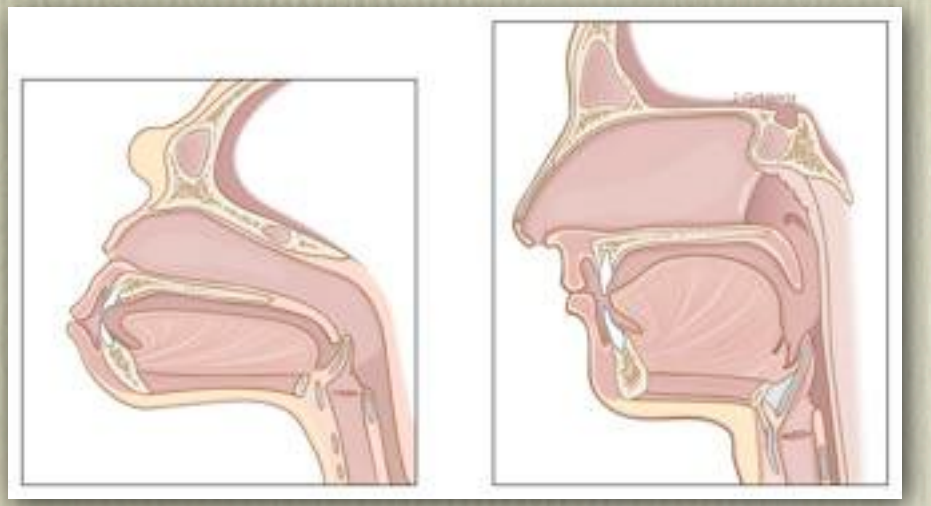
Structural diff. in the mouth and the nose



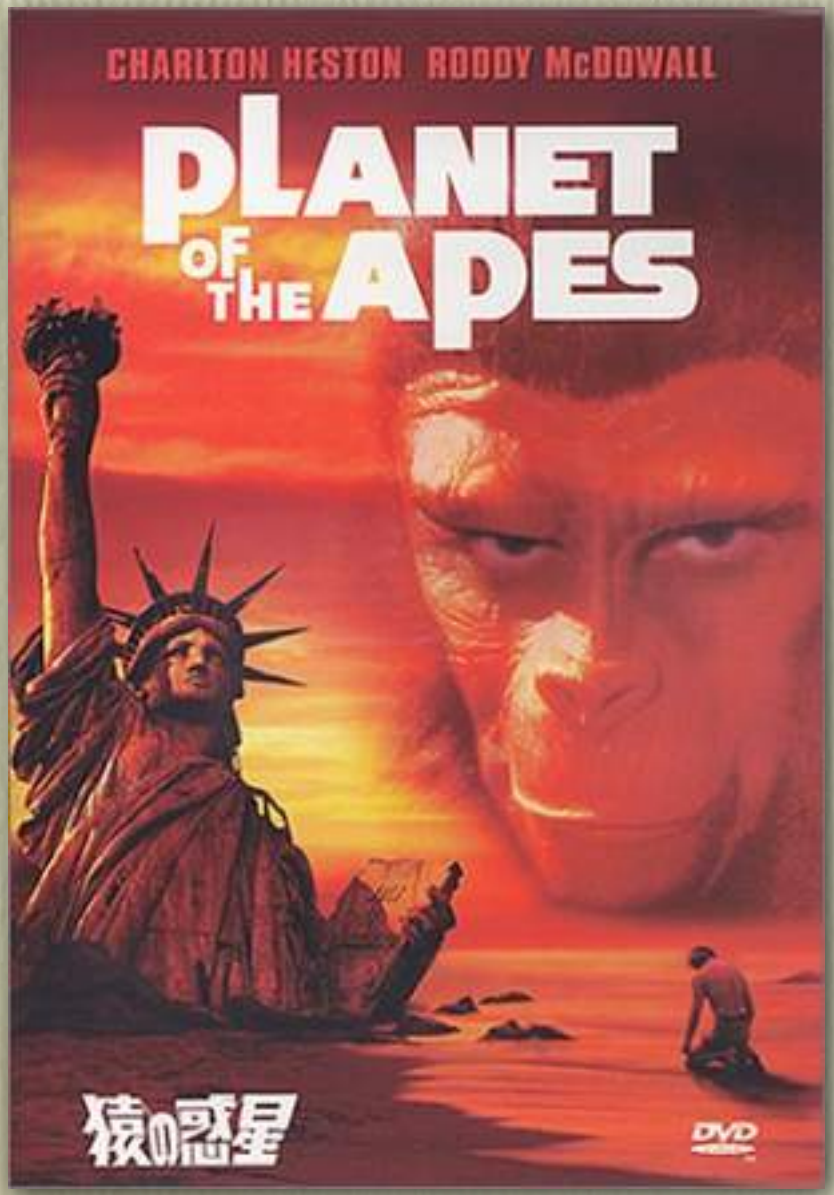
Flexibility of tongue motion

The chimp's tongue is much stiffer than the human's.

- "Morphological analyses and 3D modeling of the tongue musculature of the chimpanzee" (Takemoto'08)
  - Less capability of manipulating the shape of the tongue.



新旧「猿の惑星」

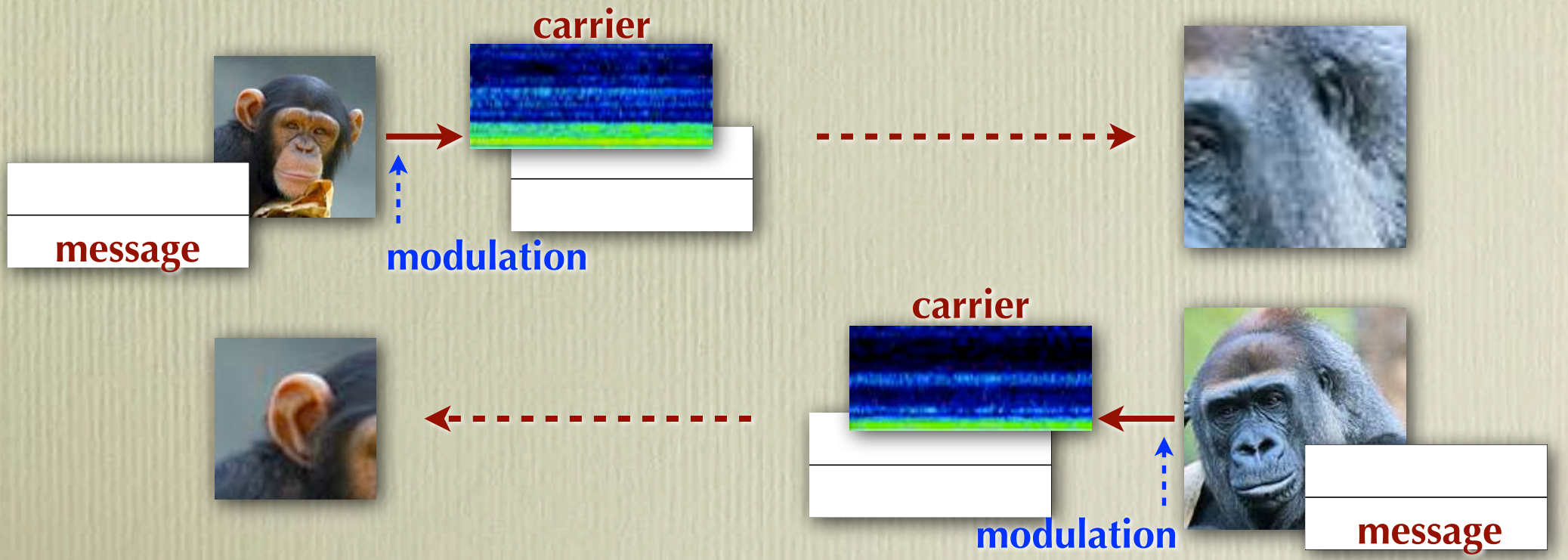

CHARLTON HESTON RODDY McDOWALL
PLANET OF THE APES
猿の惑星
DVD


DVD
+
+
DIGITAL COPY
猿の惑星
創世記
ジェネシス
RISE OF THE PLANET OF THE APES

Q1: Does the ape have a good modulator?

Morphological characteristics of the ape's tongue

- Two (almost) independent tracts [Hayama'99]
  - One is from the nose to the lung for breathing.
  - The other is from the mouth to the stomach for eating.
- Much lower ability of deforming the tongue shape [Takemoto'08]
  - The chimp's tongue is stiffer than the human's.

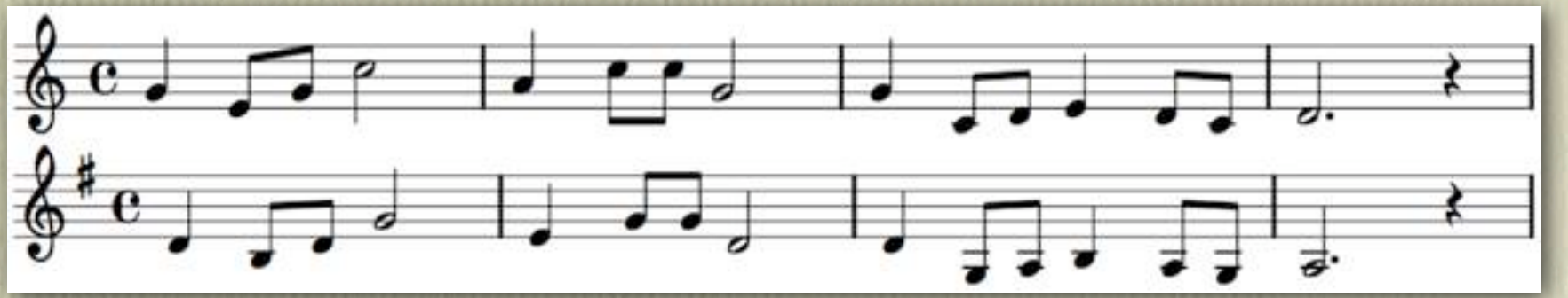


The nature's solution for static bias?

How old is the invariant perception in evolution? [Hauser'03]

1

2

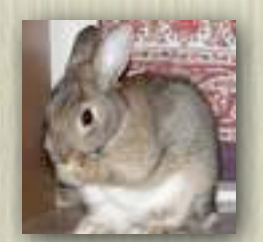
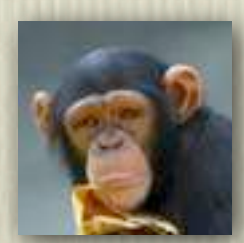


1 = 2

At least, frequency (pitch) demodulation seems difficult.






Language acquisition through vocal imitation

VI = children's active imitation of parents' utterances

- Language acquisition is based on vocal imitation [Jusczyk'00].
- VI is very rate in animals. No other primate does VI [Gruhn'06].
- Only small birds, whales, and dolphins do VI [Okanoya'08].

A's VI = acoustic imitation but H's VI ≠ acoustic = ??

- Acoustic imitation performed by myna birds [Miyamoto'95]
  - They imitate the sounds of cars, doors, dogs, cats as well as human voices.
  - Hearing a very good myna bird say something, one can guess its owner.
- Beyond-scale imitation of utterances performed by children
  - No one can guess a parent by hearing the voices of his/her child.
  - Very weird imitation from a viewpoint of animal science [Okanoya'08].

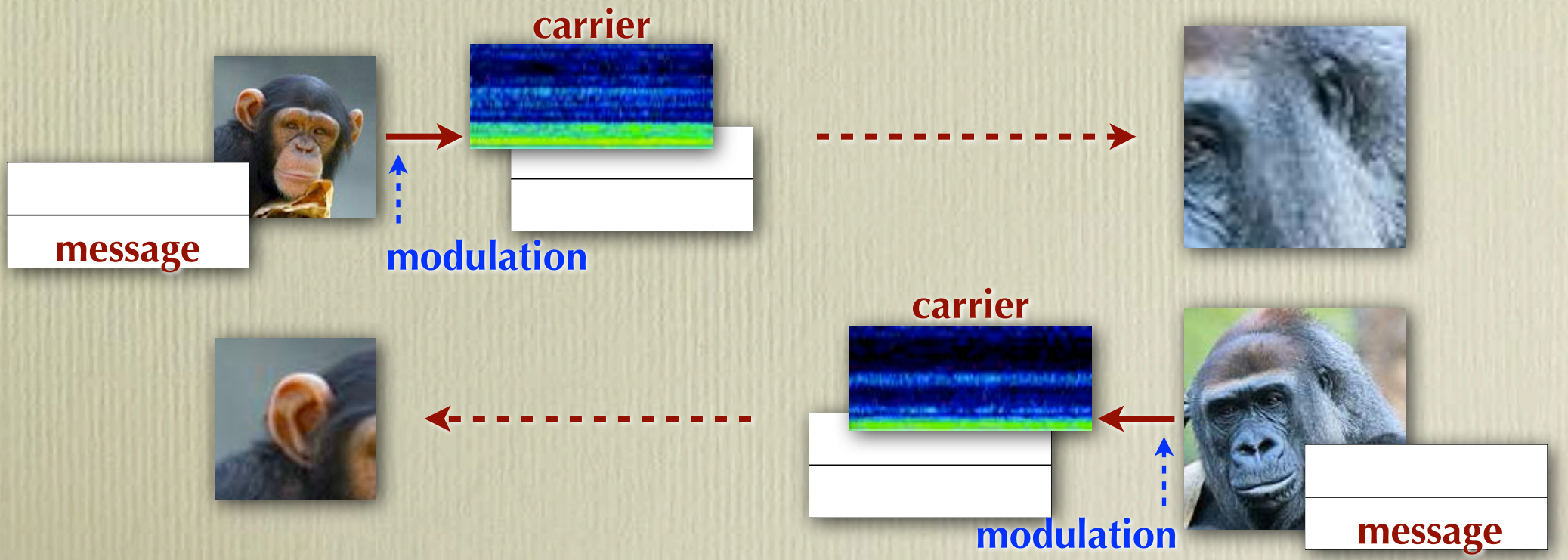





Q1: Does the ape have a good modulator?

Morphological characteristics of the ape's tongue

- Two (almost) independent tracts [Hayama'99]
  - One is from the nose to the lung for breathing.
  - The other is from the mouth to the stomach for eating.
- Much lower ability of deforming the tongue shape [Takemoto'08]
  - The chimp's tongue is stiffer than the human's.



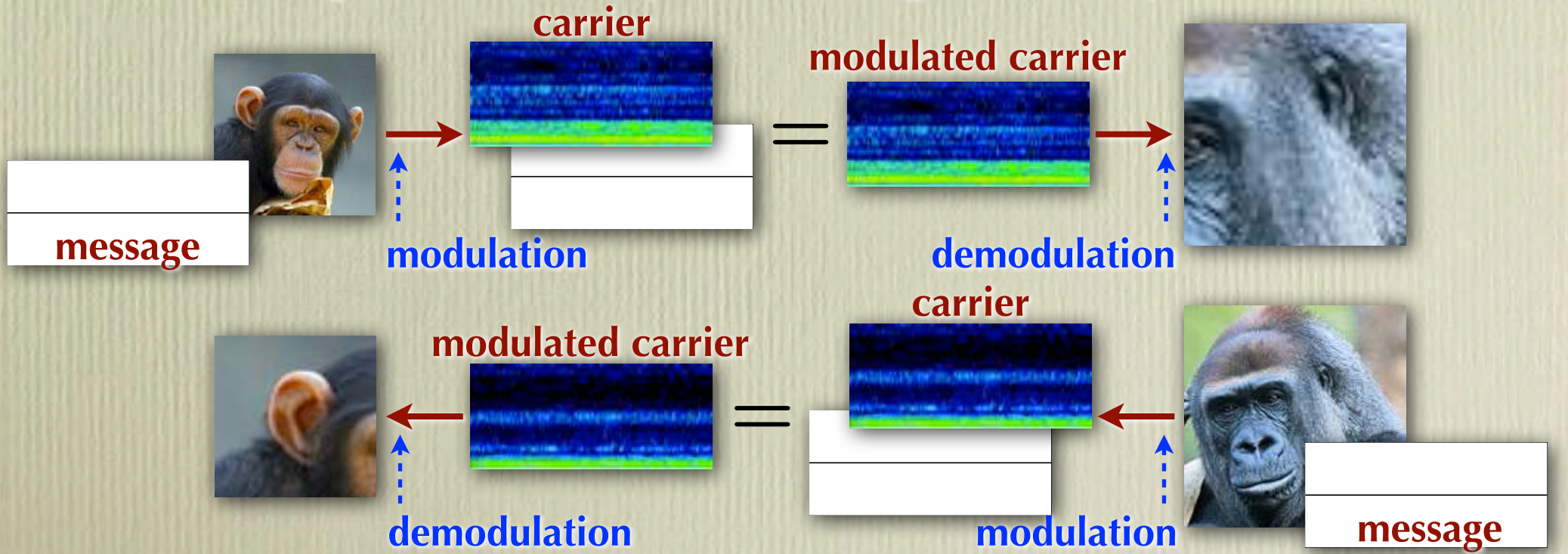
Q2: Does the ape have a good demodulator?

Cognitive difference bet. the ape and the human

- Humans can extract embedded messages in the modulated carrier.
- It seems that animals treat the modulated carrier as it is.

From the modulated carrier, what can they know?

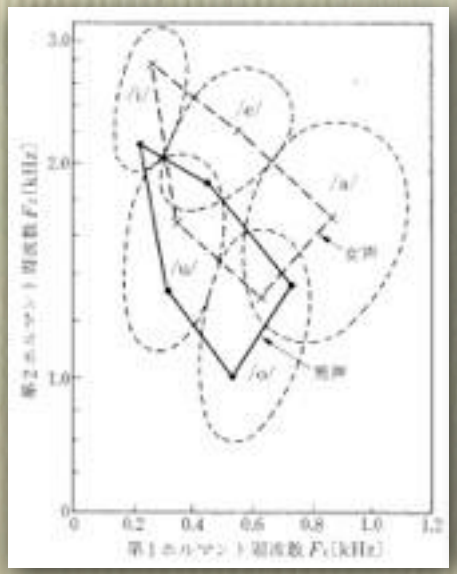
- The apes can identify individuals by hearing their voices.
  - Lower/higher formant frequencies = larger/smaller apes

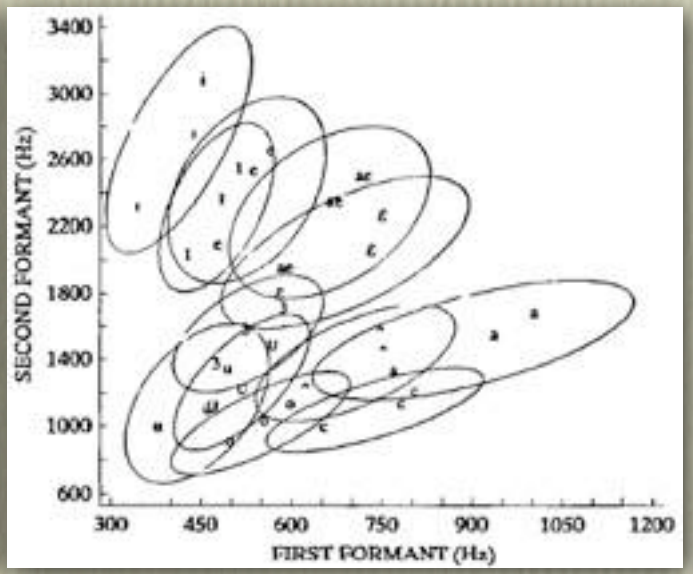


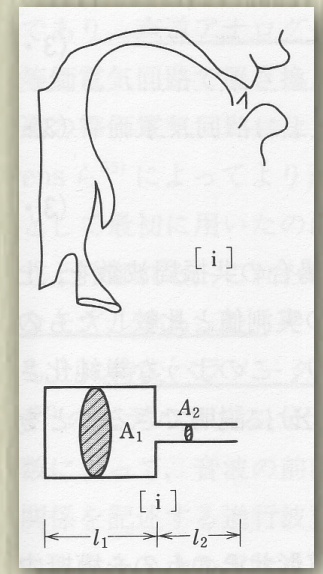
Function of the voice timbre

What is the original function of the voice timbre?

- For apes
  - The voice timbre is an acoustic correlate with the identity of apes.
- For speech scientists and engineers
  - They had started research by correlating the voice timbre with messages conveyed by speech stream such as words and phonemes.
    - Formant frequencies are treated as acoustic correlates with vowels.
  - “Speech recognition” started first, then, “speaker recognition” followed.







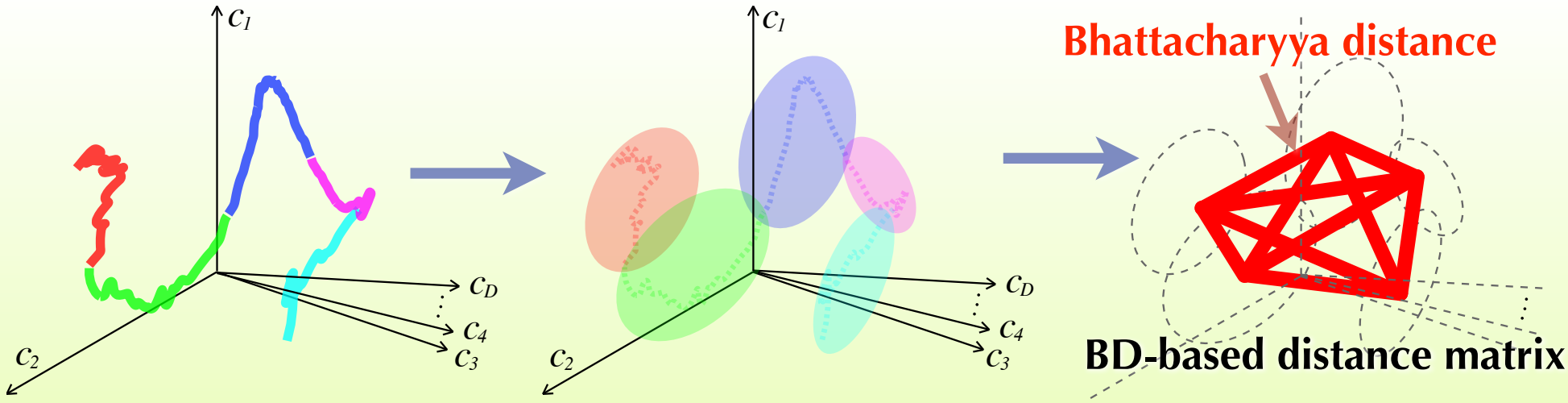
$$f_n = \frac{c}{2l_1} n$$

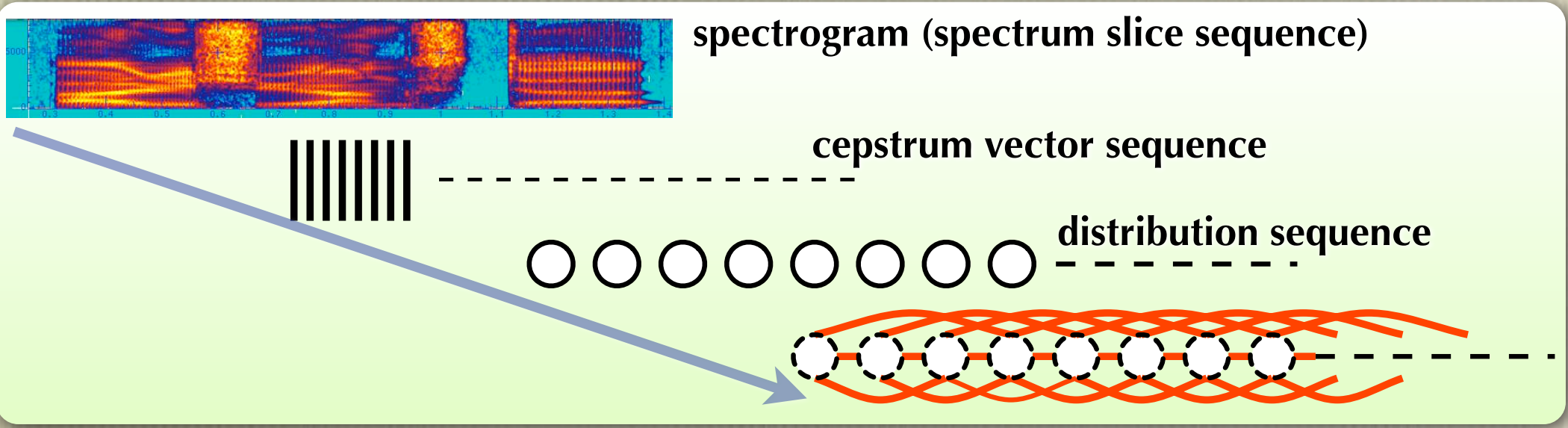
$$f_n = \frac{c}{2l_2} n$$

$$f = \frac{c}{2\pi} \left[\frac{A_2}{A_1 l_1 l_2} \right]^{1/2}$$

Invariant speech structure

Utterance to structure conversion using *f*-div. [Minematsu'06]





- An event (distribution) has to be much smaller than a phoneme.

シンボルグラウンディング問題

シンボルは如何にして生まれたか？

- シンボルは何故「意味」を持つようになったのか？
- シンボルは何故「ある対象物」と結びつくようになったのか？
- シンボルは何故「ある記憶」と結びつくようになったのか？

寄 稿

特集 ことばをとどける「声の力」

声とは、言葉とは、何か

——音声研究を通して考えること

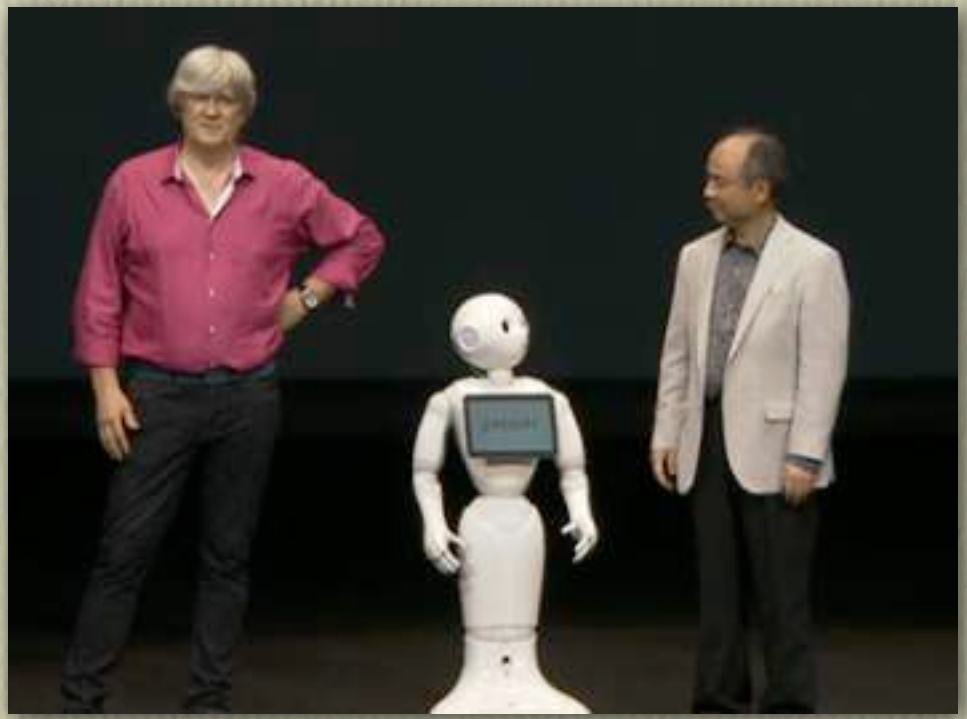
東京大学大学院工学系研究科教授

峯松 信明

声とは何か、言葉とは何か。この根源的なテーマに応えてくださったのは、音声工学の第一人者である峯松信明先生。機械に音声を認識させる・合成させる、その研究を通して対極に見えてきたものとは何でしょうか。それはヒトの持つ不思議な能力——言葉と記憶、ヒトは言葉を操作しながら、実は言葉によって記憶を操作されている——その謎に科学の目で迫ります。

プロフィール／みねまつ・のぶあき
1990年 東京大学工学部卒業、95年東京大学大学院工学系研究科にて博士（工学）を取得。95年より豊橋技術科学大学に勤務し、2000年より東京大学に戻る。現在、東京大学大学院工学系研究科電気系工学専攻教授。音声科学から音声工学に至るまで、幅広い観点から音声コミュニケーションに関する研究に従事。特に音声技術を使った語学教育に関する造詣が深く、2009年よりOJADの開発を手がけている。

What is the goal of speech engineering?



計算できる馬

賢馬ハンスから学べること

何が欠けているのか？

二つの軸

- 発達
  - 我々は成長の中でどのように言語を身に付けるのか？
- 進化
  - 我々は進化の中でどのように言語を身に付けたのか？
- この二軸を真っ正面に見据えて技術開発しないと・・・・
  - それは，言葉を操るように見せかけるシステムとなる，のでは？

高校生のためのオープンキャンパスにて

言葉が分かるコンピュータってどんなコンピュータ？

東大で言葉の研究をする工学系教員から高校生への素朴な問いかけ

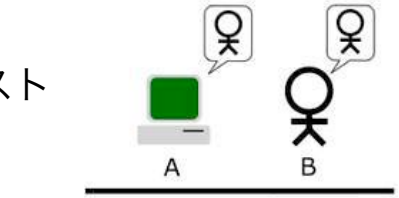
■Siri，喋ってコンシェル，IBM Watson，彼らは「言葉が分かる」コンピュータなのか？

「ニューヨークは今何時？」「8月6日午後10時です」
「清水寺の舞台の高さは？」「約13メートルです」
「ソーダ瓶の回転が止まった時に，瓶の口の前にいる人は唇を突き出すゲームは？」「Spin-the-bottleです」
彼らは話された／書かれた内容を理解して，吟味して，返答しているように見える。
では，彼らは本当に**「言葉が分かる」**のか，それとも**「言葉が分かったように見せかけている」**だけなのか？

このポスターは**「言葉が分かる」**とはどういうことなのか，高校生の皆さんにちょっと深く考えてもらいたくて作りました。上の問いに対して先人達はどのように考えてきたのか，を紹介します。もしかしたら，本当に言葉が分かるコンピュータを作ることになるのは，数年後，いや数十年後の貴方，かもしれません。

■「チューリング・テスト」って知ってますか？

数学者アラン・チューリングが考案した**「ある機械が知的であるかどうか」**を判定するテスト



人間の判定者Cが，隔離された相手A，Bと通常の言語で会話する。A，Bは一方が機械，他方が人間である。会話の後Cはどちらが人間／機械なのかを当る。その区別が困難であれば，この機械はテストに合格，つまり，知的であると判定する。

今でも「人工知能」研究でしばしば利用される判定基準である。

■「中国語の部屋」って知ってますか？

高校生のためのオープンキャンパスにて

■「中国語の部屋」って知ってますか？

チューリングテストに対して哲学者ジョン・サールが問うた鋭い突っ込み（思考実験）

ある小部屋にアルファベットしか理解できない人を閉じこめておく。この部屋には外部と紙切れのやりとりをする穴が一つ空いている。この穴を通してこの人に一枚の紙切れが差し入れられる。そこには漢字で何か書いてあるが，彼には単なる記号列でしかない。彼の仕事はこの記号列に対して，新たな記号列を書き加えて外に返すことである。どういう記号列を書き加えればよいのかは，一冊のマニュアルに書いてある。例えば「★△◎▽☆□」とあれば，「■@◎▽」と書き加えて外に出せ，のように。

部屋の外で紙切れを観測している人にすれば「中国語が分かる人が内部にいる」と考えるだろう。部屋にいるのは漢字が全然理解できない人なのに。

■XXするように見せかけている例というのは，結構沢山あるのかも・・・

プラネタリウム：あれは基本的に天動説に基づいて星を動かしています。座席は動きませんから。でも，星の見た目の動きを再現するという目的であれば，天動説も地動説も結果は殆ど変わりませんよね。

賢馬ハンス：20世紀初頭，ドイツで有名になった「計算できる」馬。後に科学的手法によりトリックが判明。

DaiGo：21世紀初頭，日本のテレビ業界を賑わしているメンタリスト。彼の場合は「トリックがあります」と自分で明言してますけど。

見た目を上手に作り込むのか，中のメカニズムにまでこだわるのか？

■結局，何ができれば「言語が分かる」コンピュータなのか，その定義が難しいのですよ。

高校生のためのオープンキャンパスにて

■XXするように見せかけている例というのは，結構沢山あるのかも・・・

プラネタリウム：あれは基本的に天動説に基づいて星を動かしています。座席は動きませんから。でも，星の見た目の動きを再現するという目的であれば，天動説も地動説も結果は殆ど変わりませんよね。

賢馬ハンス：20世紀初頭，ドイツで有名になった「計算できる」馬。後に科学的手法によりトリックが判明。

DaiGo：21世紀初頭，日本のテレビ業界を賑わしているメンタリスト。彼の場合は「トリックがあります」と自分で明言してますけど。

見た目を上手に作り込むのか，中のメカニズムにまでこだわるのか？

■結局，何ができれば「言語が分かる」コンピュータなのか，その定義が難しいのですよ。

「言語が分かる」コンピュータを実現するための必要十分条件の定義が難しい。できるのは，必要条件を洗い出すことだけなのかもしれない。で，**どの**必要条件に着目し，技術として実装するのか，それは各研究者のこだわりとなって，研究戦略に現れるのだと思います。さてさて，貴方が**「言語が分かる」**コンピュータを作ろうとしたら，どんなコンピュータを作りますか？　**貴方自身の答え**を，この部屋で見つけてみて下さい。