Psychoinformatics & Neuroinformatics

Week 14

Audio, Speech,
& Language Processing

by Tsung-Ren (Tren) Huang 黃從仁

Topics for today

Audio Processing

Extraction of sound features

Speech Processing

Speech2Text & Text2Speech

Language Processing

Making (voice) chatbots

Topics for today

Audio Processing

Extraction of sound features

Speech Processing

Speech2Text & Text2Speech

Language Processing

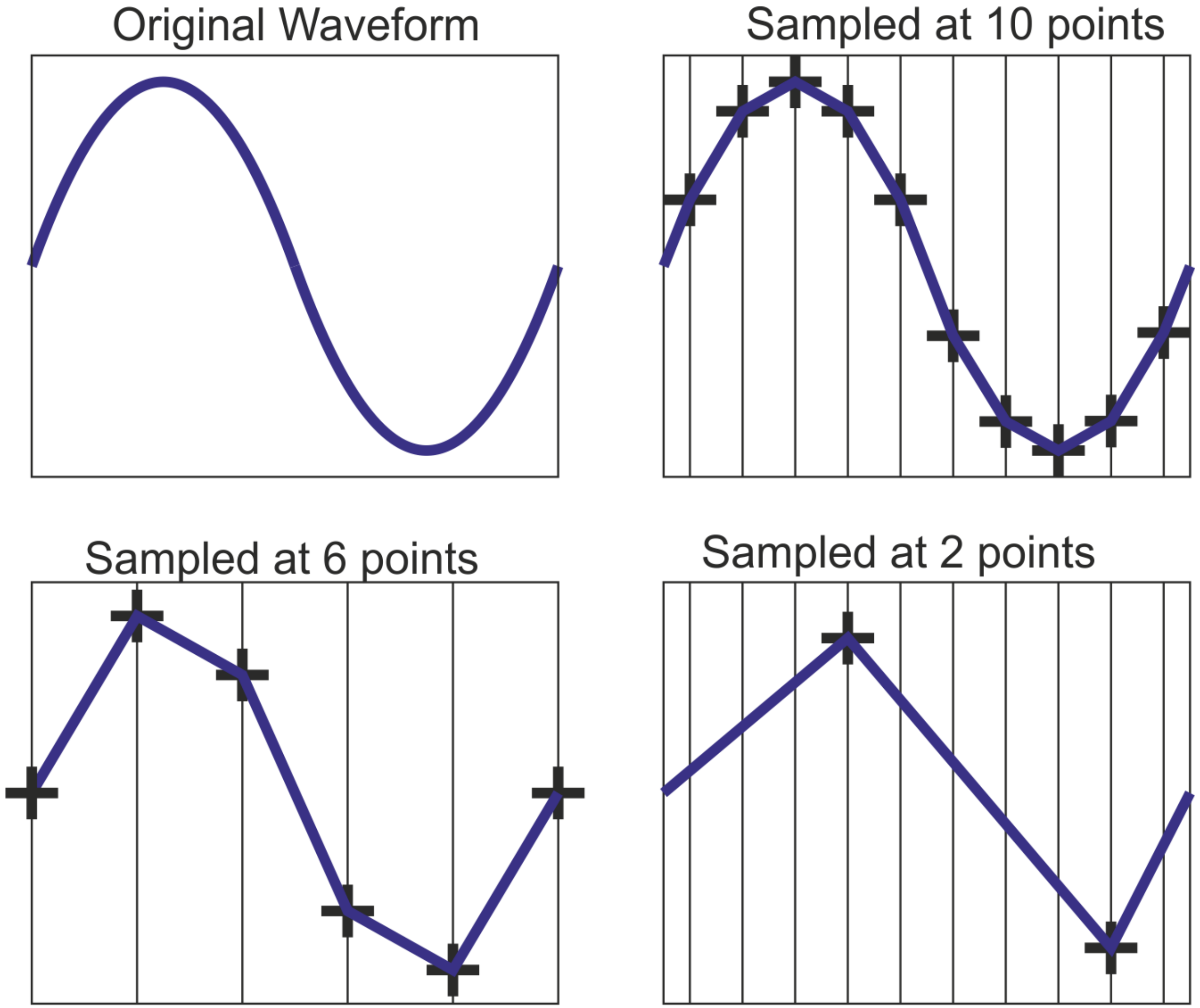
Making (voice) chatbots

Case Study: Paralinguistic Features

Both visual and audio signals are informative of emotions:

Digital Signal Processing (1/3)

Original sound signals are in the time domain:

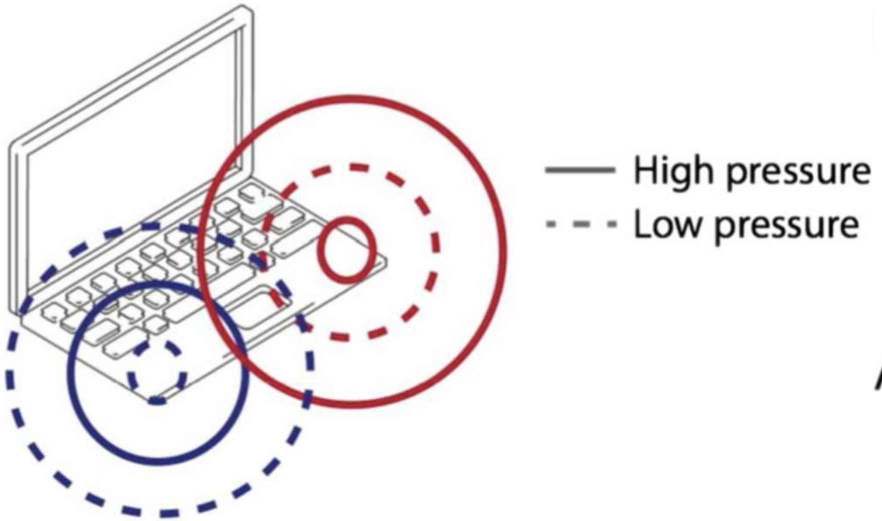


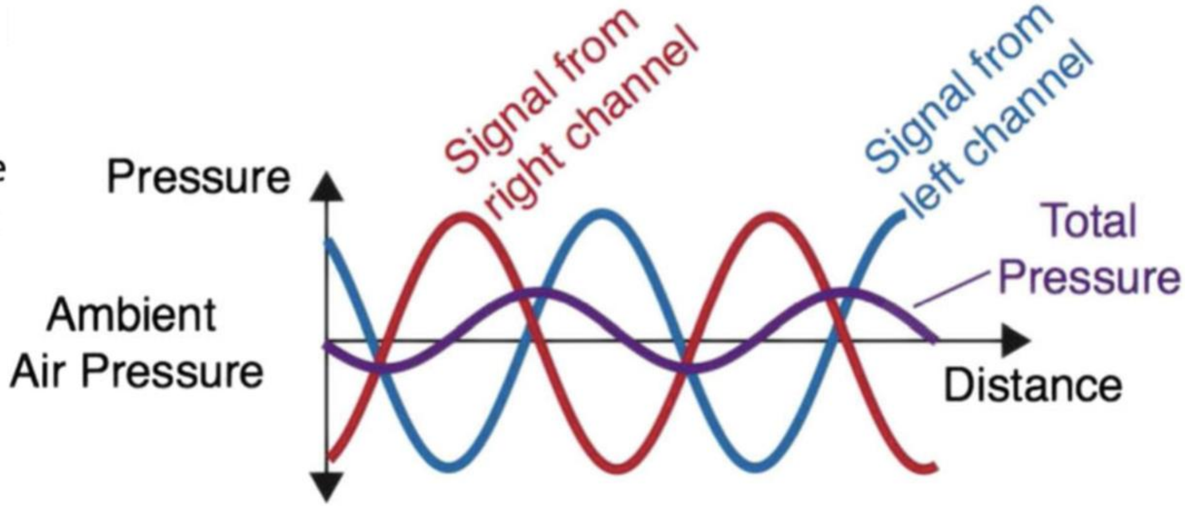
Digital Signal Processing (2/3)

Two-channel out-out-phase audio signals will cancel each other out when played by speakers:

Headphone screening to facilitate web-based auditory experiments

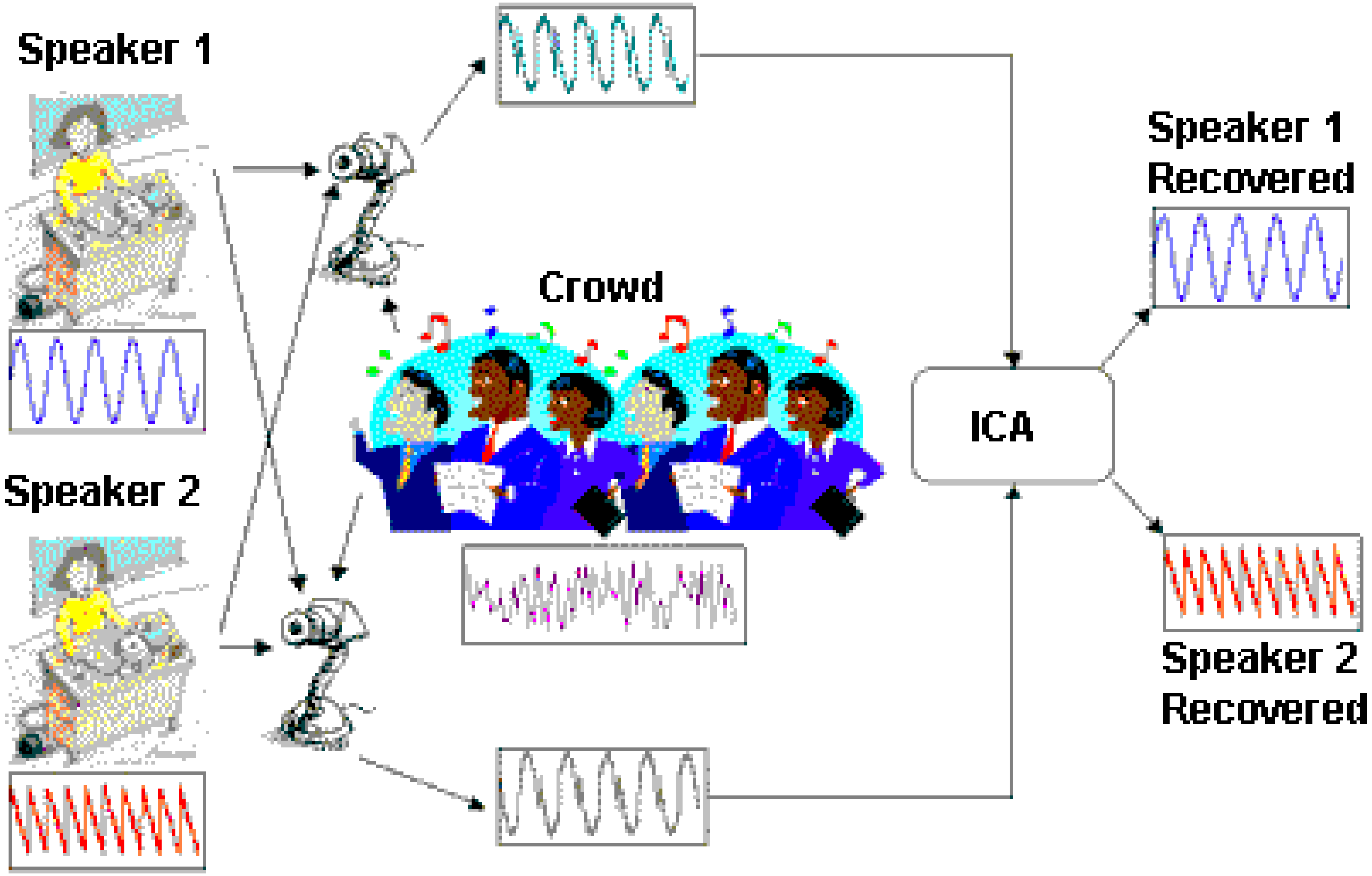
Kevin J. P. Woods[1,2] · Max H. Siegel[1] · James Traer[1] · Josh H. McDermott[1,2]





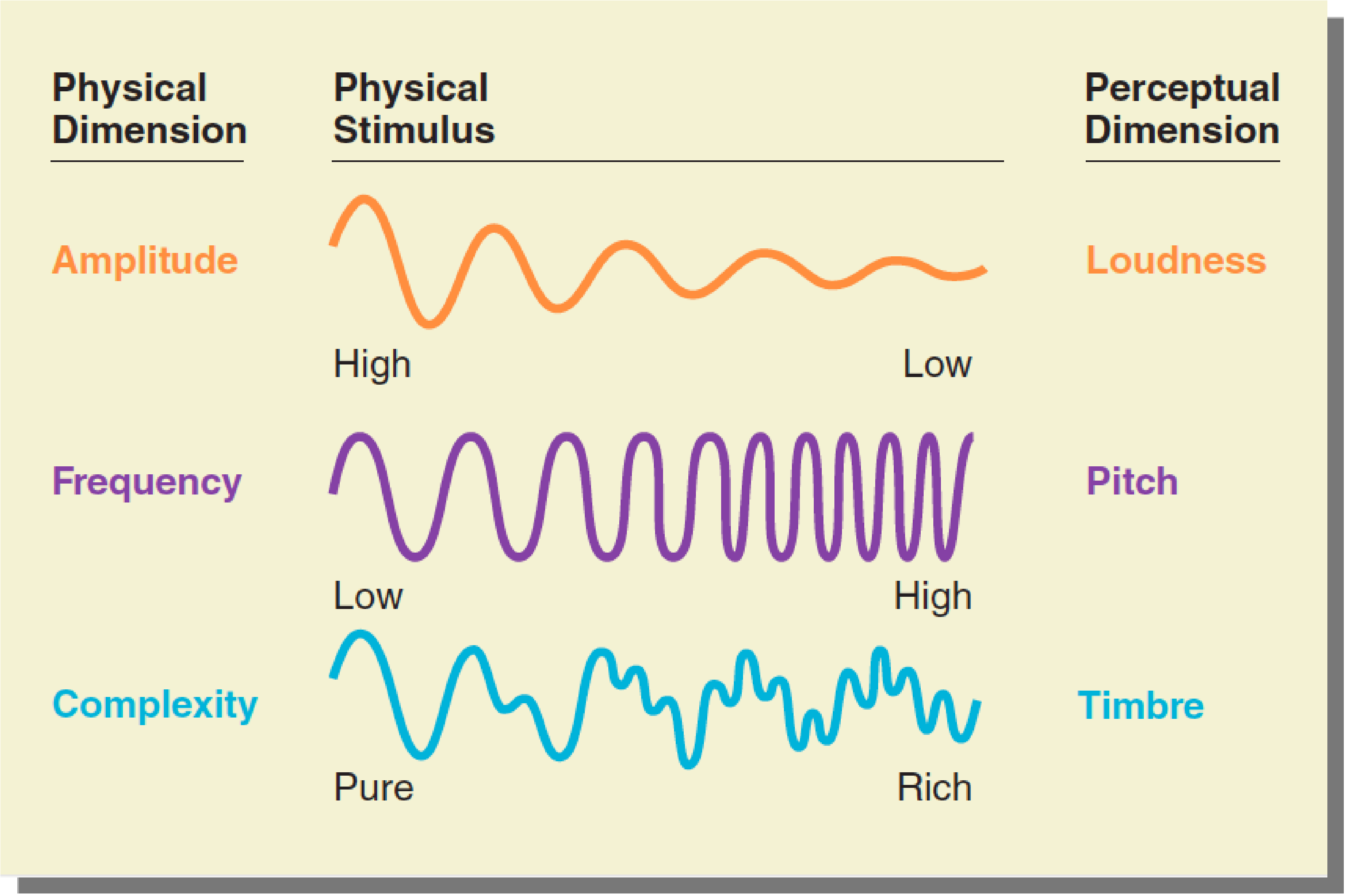
Digital Signal Processing (3/3)

ICA can be used for speaker diarization:



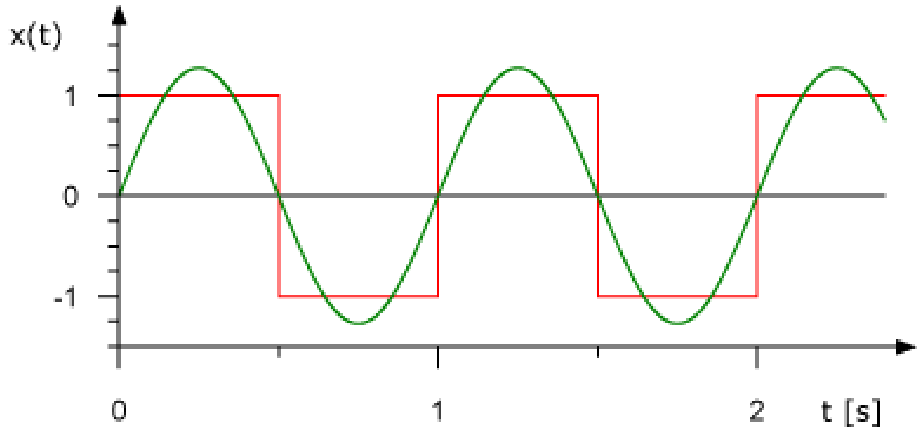
Physical vs. Perceptual Dimensions

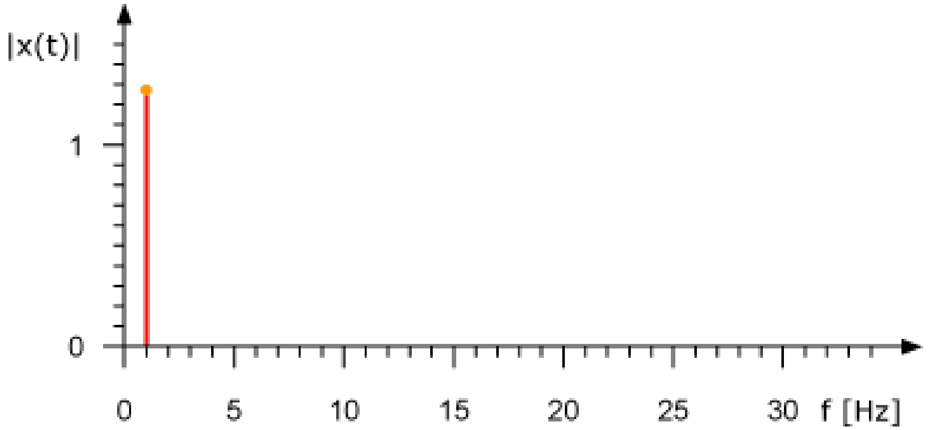
Instrument : Timbre :: Person : Voiceprint



Fourier Transform: Frequency Domain

$$y(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \left[a_k \cos(2\pi k f_0 t) - b_k \sin(2\pi k f_0 t) \right]$$

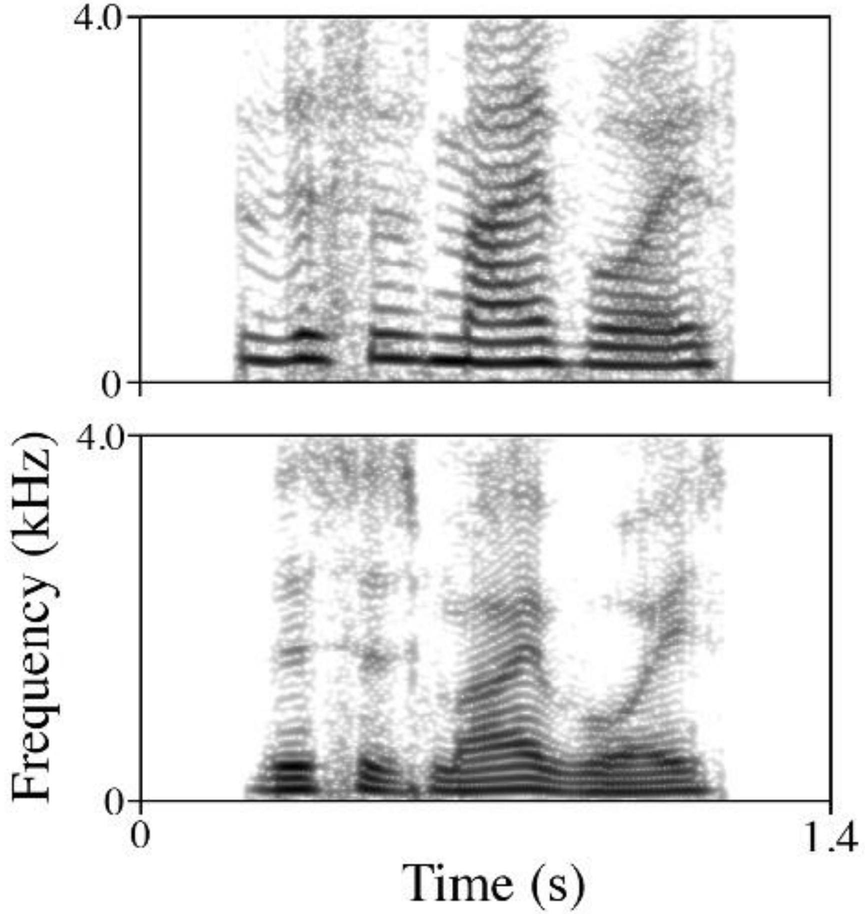


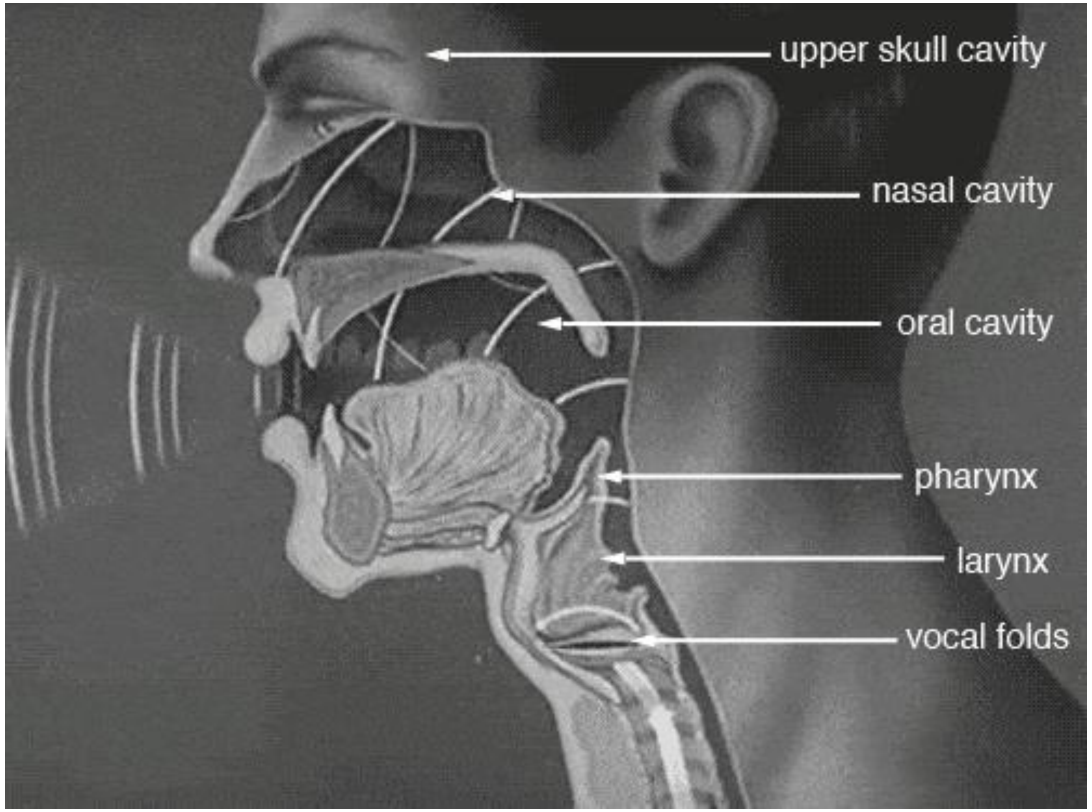


Short-Time FT (STFT): Spectrogram

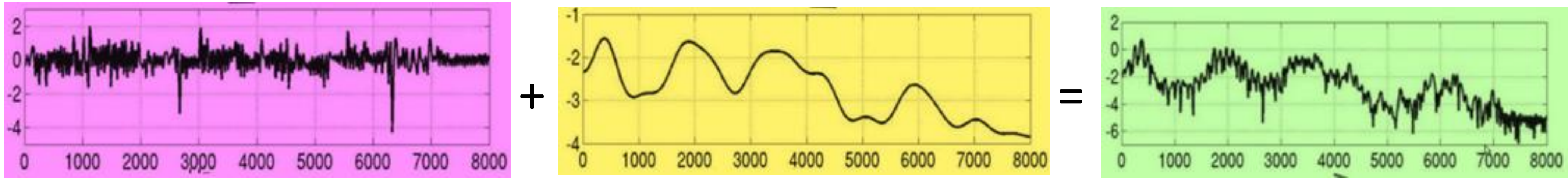
A series of Fourier Transforms for (sliding) time windows

"This is my voice"





Cepstrum: $C_p = |FT^{-1}\{\log(|FT\{s(t)\}|^2)\}|^2$

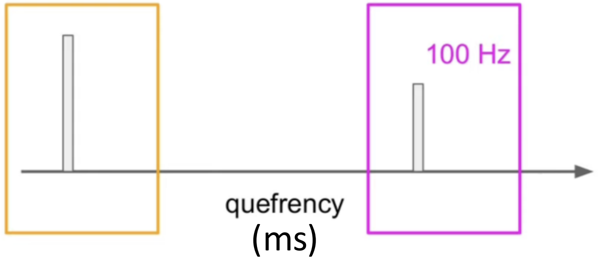


Time domain: $g(t)*v(t)=s(t)$

Frequency domain: $G(f)\cdot V(f)=S(f)$

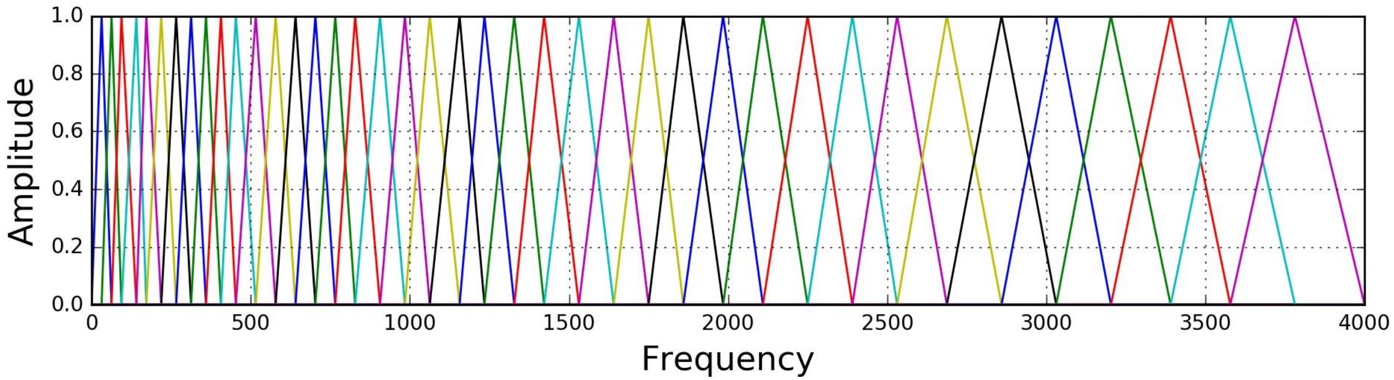
Log-frequency: $\log(G(f))+\log(V(f))=\log(S(f))$

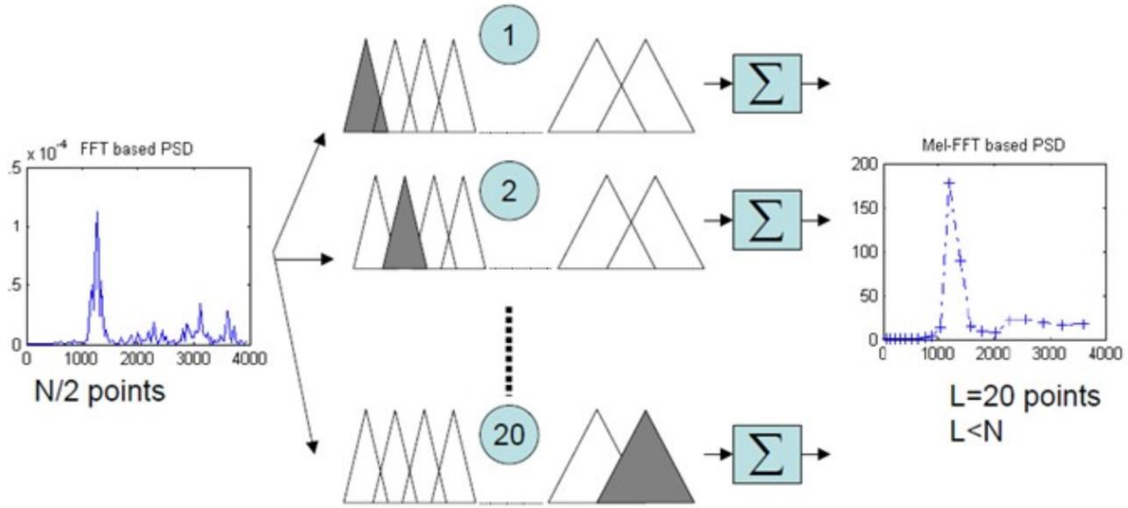
FT again to get C_p



Mel Filters for nonuniform sampling

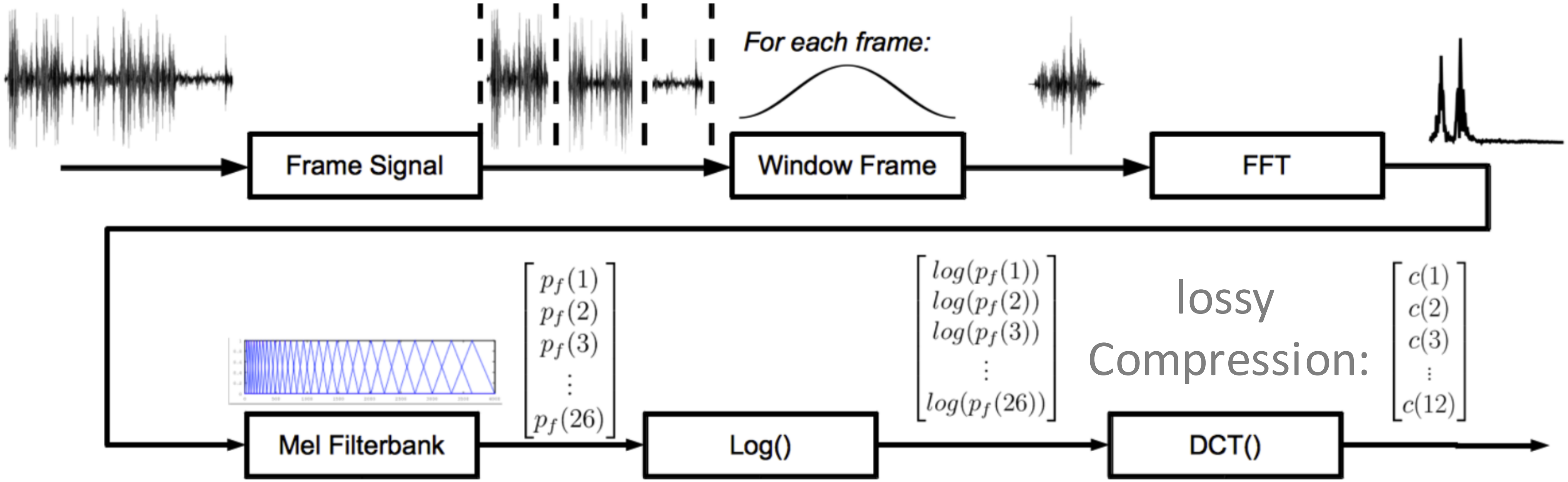
Human can only detect sounds of 20~20,000Hz, which covers male speech (85~155Hz) & female speech (165~255Hz).

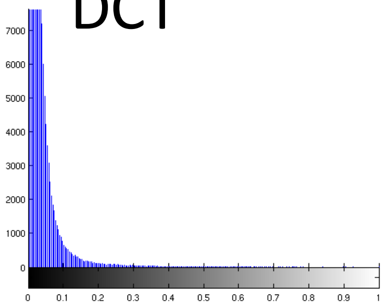


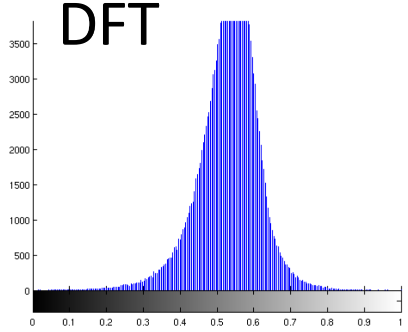


Mel-scale Frequency Cepstral Coefficients

DCT, a real-valued FT, is used to decorrelate/compress log(p)







Various Feature Sets

openSMILE:) by audEERING™ can help extract various features!

2.5 Default feature sets

For common tasks from the Music Information Retrieval and Speech Processing fields we provide some example configuration files in the `config/` directory for the following frequently used feature sets. These also contain the baseline acoustic feature sets of the 2009–2013 INTERSPEECH challenges on affect and paralinguistics:

- Chroma features for key and chord recognition
- MFCC for speech recognition
- PLP for speech recognition
- Prosody (Pitch and loudness)
- The INTERSPEECH 2009 Emotion Challenge feature set
- The INTERSPEECH 2010 Paralinguistic Challenge feature set
- The INTERSPEECH 2011 Speaker State Challenge feature set
- The INTERSPEECH 2012 Speaker Trait Challenge feature set
- The INTERSPEECH 2013 ComParE feature set
- The MediaEval 2012 TUM feature set for violent scenes detection.

Acoustic LLDs	
Low-level Descriptors (LLDs)	Type
zero-crossing rate, log energy, probability of voicing, F_0	prosodic
MFCC 0-12, spectral flux, spectral centroid, max, min, spectral bands 0-4 (0-9KHz), spectral roll-off (0.25, 0.5, 0.75, 0.9)	spectral
Functionals applied to LLDs/ΔLLDs/ΔΔLLDs	
position of min/max, range, max – arithmetic mean, arithmetic mean – min	extremes
linear regression slope, offset, error, centroid, quadratic error, quadratic regression a, b offset, linear error, quadratic error (contour & quadratic regression)	regression
percentile range (25%, 50%, 75%), 3 inter-quartile ranges (25% - 50%, 50%-75%, 25%-75%)	percentiles
mean value of peaks, distance between peaks, mean value of peaks – arithmetic mean	peaks
arithmetic means, absolute value of arithmetic mean (original, non-zero values), quadratic mean (original, non-zero values), geometric mean (absolute values of non-zero values), number of non-zero values	means
relative duration LLD above 25%, 50%, 75%, 95% range, relative duration LLD is rising/falling, relative duration LLD has left/right curvature	temporal

Topics for today

Audio Processing

Extraction of sound features

Speech Processing

Speech2Text & Text2Speech

Language Processing

Making (voice) chatbots

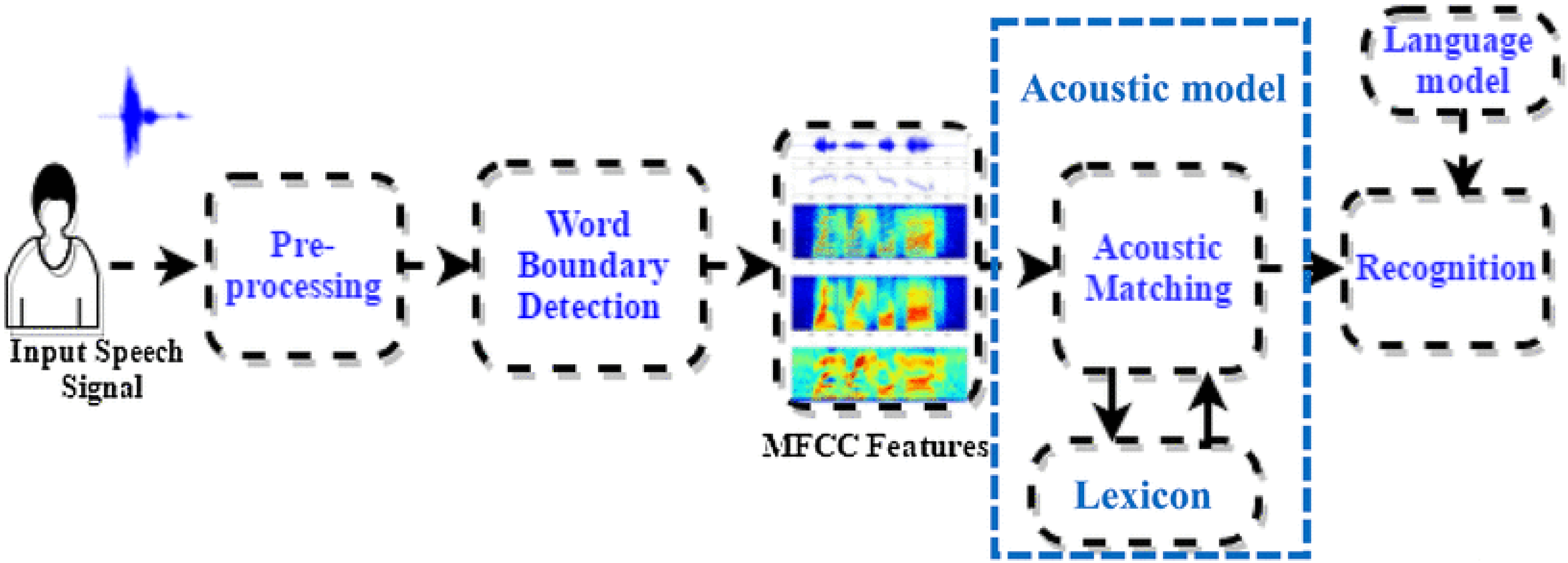
Case Study: Number of Spoken Words

Mehl et al., 2007, *Science*

Sample	Year	Location	Duration	Age range (years)	Sample size (N): Women	Sample size (N): Men	Estimated average number (SD) of words spoken per day: Women	Estimated average number (SD) of words spoken per day: Men
1	2004	USA	7 days	18–29	56	56	18,443 (7460)	16,576 (7871)
2	2003	USA	4 days	17–23	42	37	14,297 (6441)	14,060 (9065)
3	2003	Mexico	4 days	17–25	31	20	14,704 (6215)	15,022 (7864)
4	2001	USA	2 days	17–22	47	49	16,177 (7520)	16,569 (9108)
5	2001	USA	10 days	18–26	7	4	15,761 (8985)	24,051 (10,211)
6	1998	USA	4 days	17–23	27	20	16,496 (7914)	12,867 (8343)
					Weighted average		16,215 (7301)	15,669 (8633)

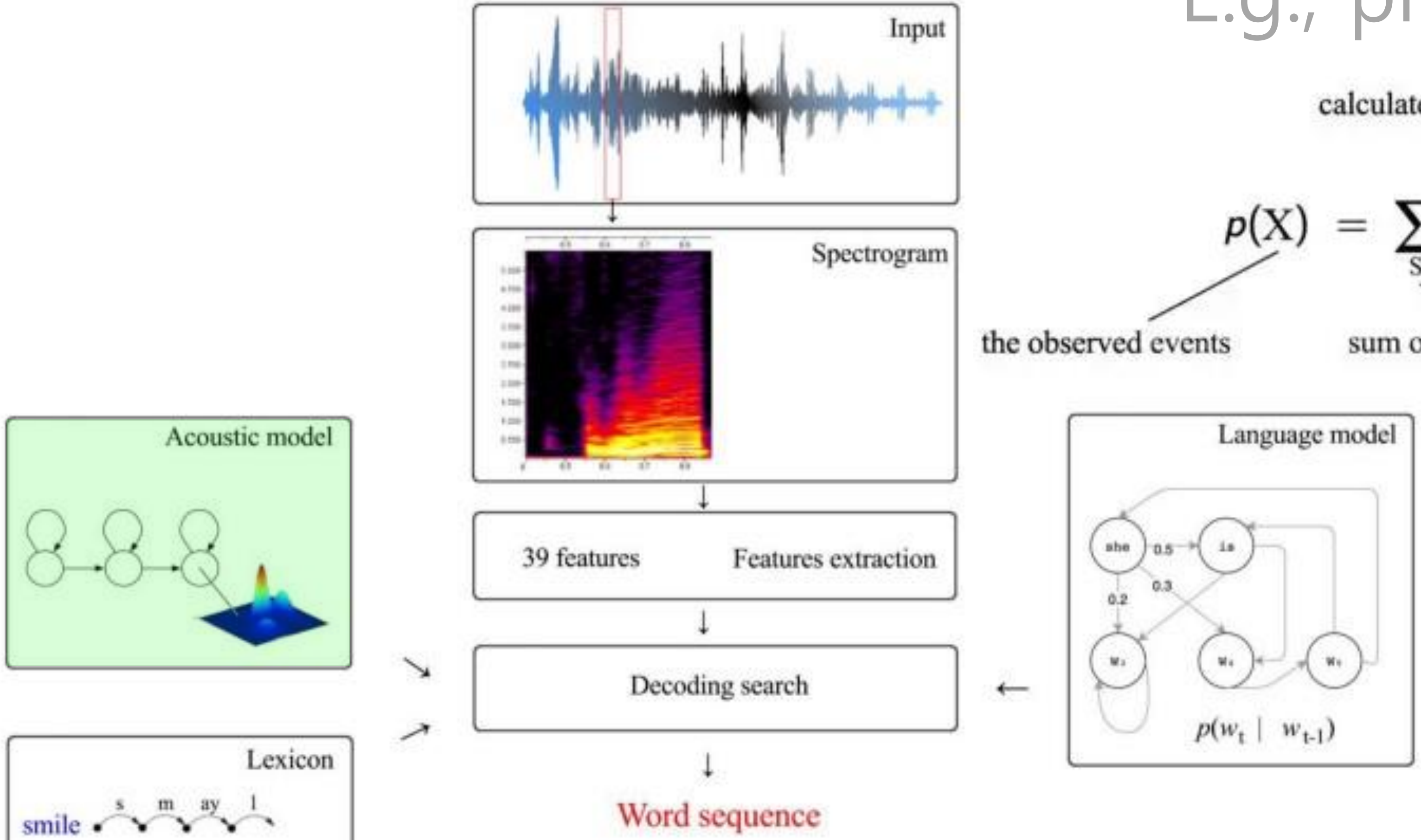
Speech Recognition (1/4): Matching

Bottom-up: Matching each input to feature templates of words



Top-down: Contexts help to disambiguate (e.g., close vs. clothes)

Speech Recognition (2/4): HMM



$$\mathbf{W}^* = \arg\max_{\mathbf{W}} P(\mathbf{W} \mid \mathbf{X})$$

$$\mathbf{W}^* = \arg\max_{\mathbf{W}} \; p(\mathbf{X} \mid \mathbf{W}) \;\; P(\mathbf{W})$$

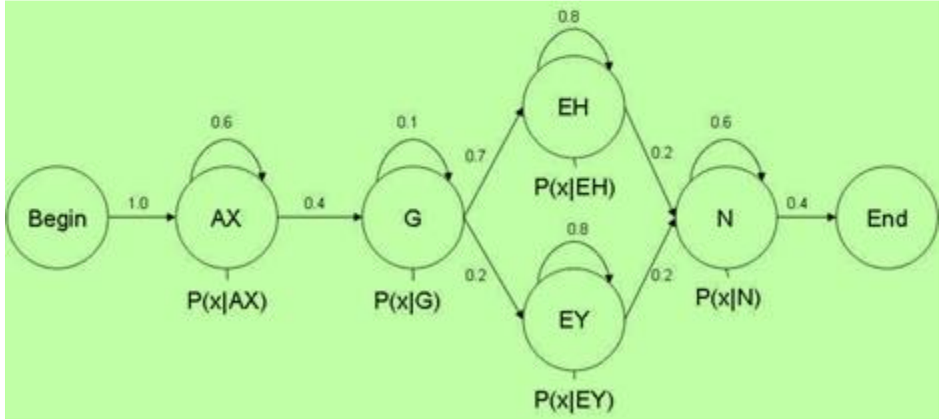
word sequence acoustic model language model

E.g., present, desert, IKEA, etc.

calculated from emission probability calculated from transition probability

$$p(\mathrm{X}) = \sum_{\mathrm{S}} p(\mathrm{X}, \mathrm{S}) = \sum_{\mathrm{S}} p(\mathrm{X} \mid \mathrm{S})\, p(\mathrm{S})$$

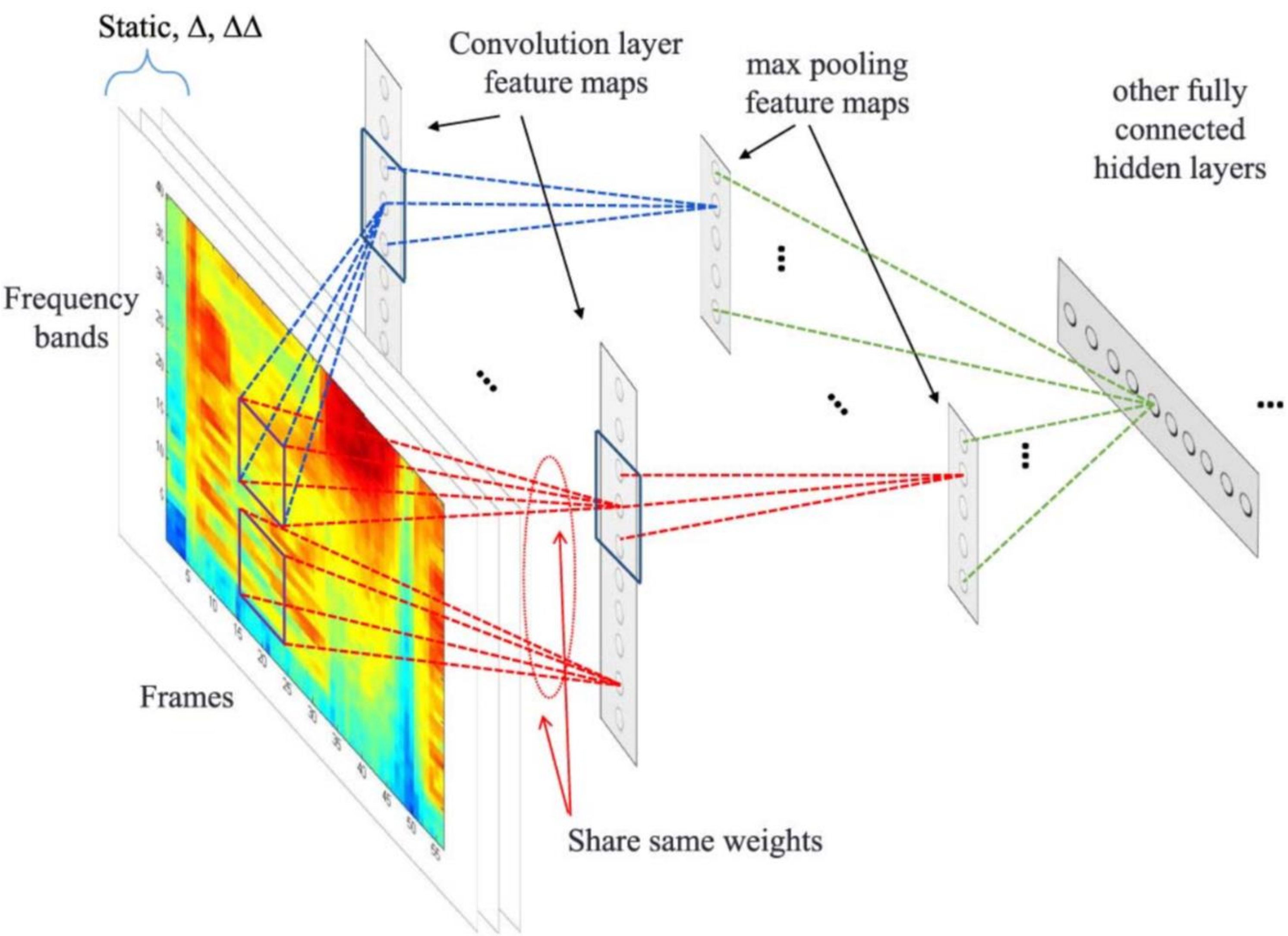
the observed events sum over all possible time sequences of internal states



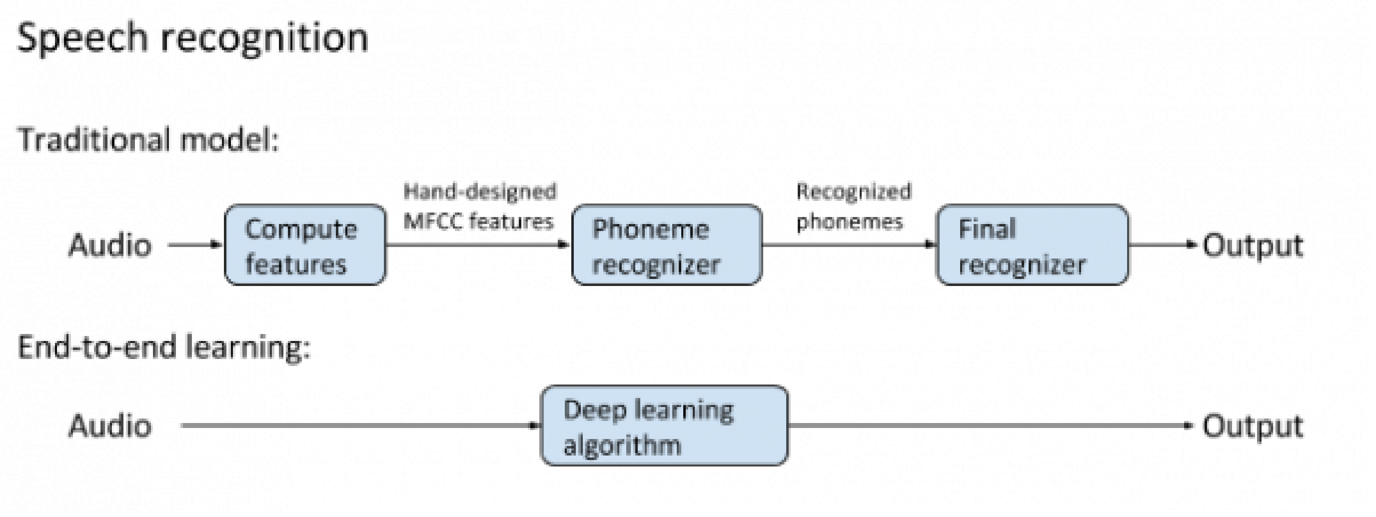
Hidden Markov Model for the word "again"

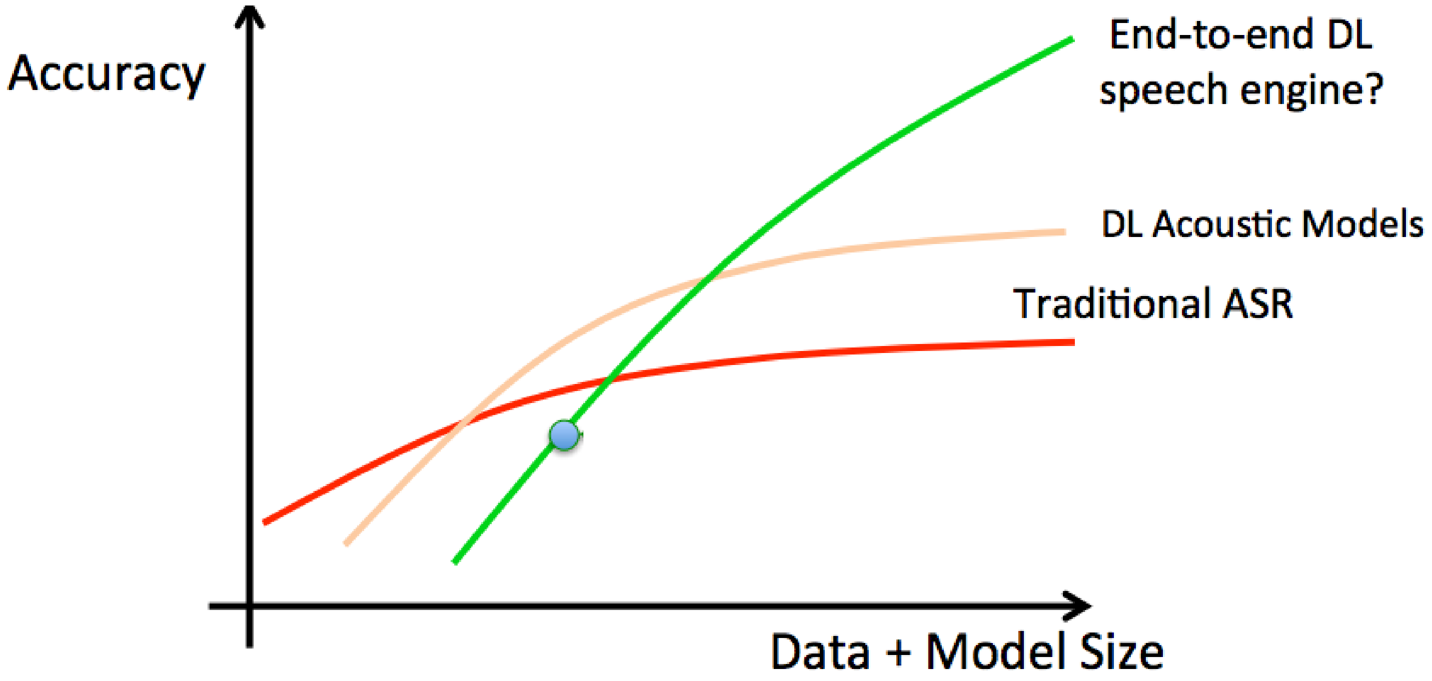
Speech Recognition (3/4): CNN

Use CNN to recognize spectrogram/cepstrogram as an image



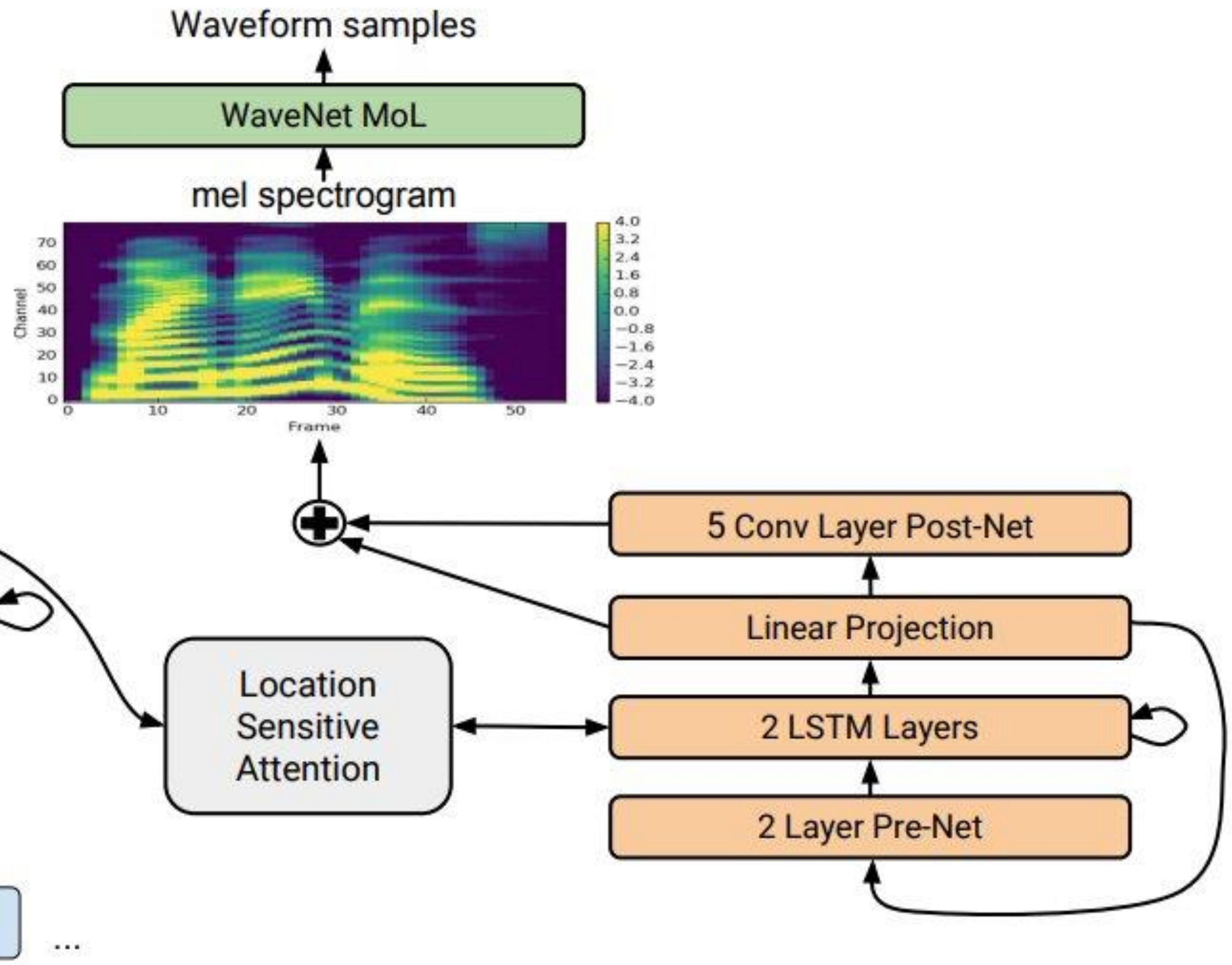
Speech Recognition (4/4): Summary





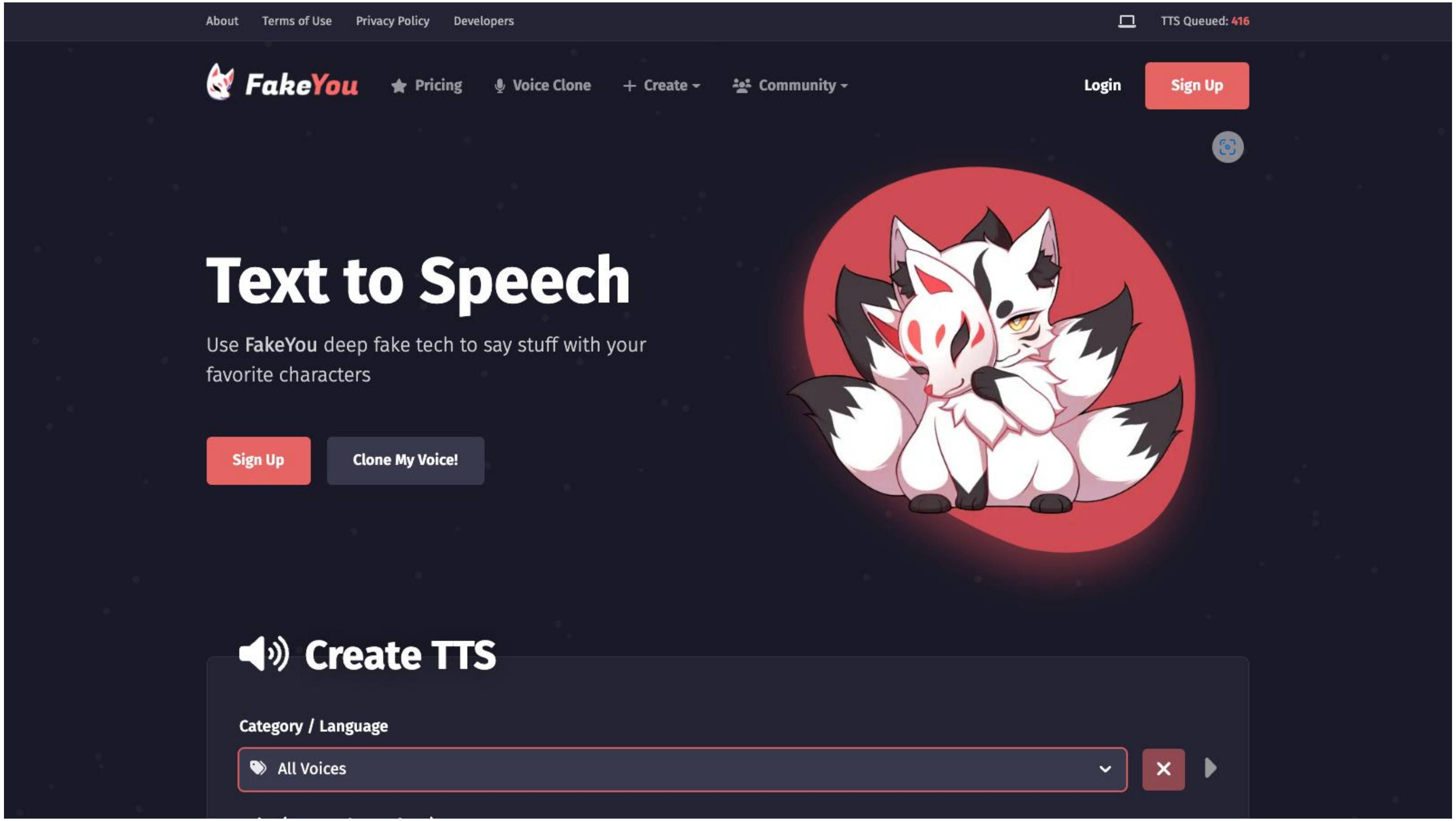
Speech Synthesize (1/2): Concepts

Namely, text to speech:



Speech Synthesize (2/2): Apps

There are many apps available for text to speech



Topics for today

Audio Processing

Extraction of sound features

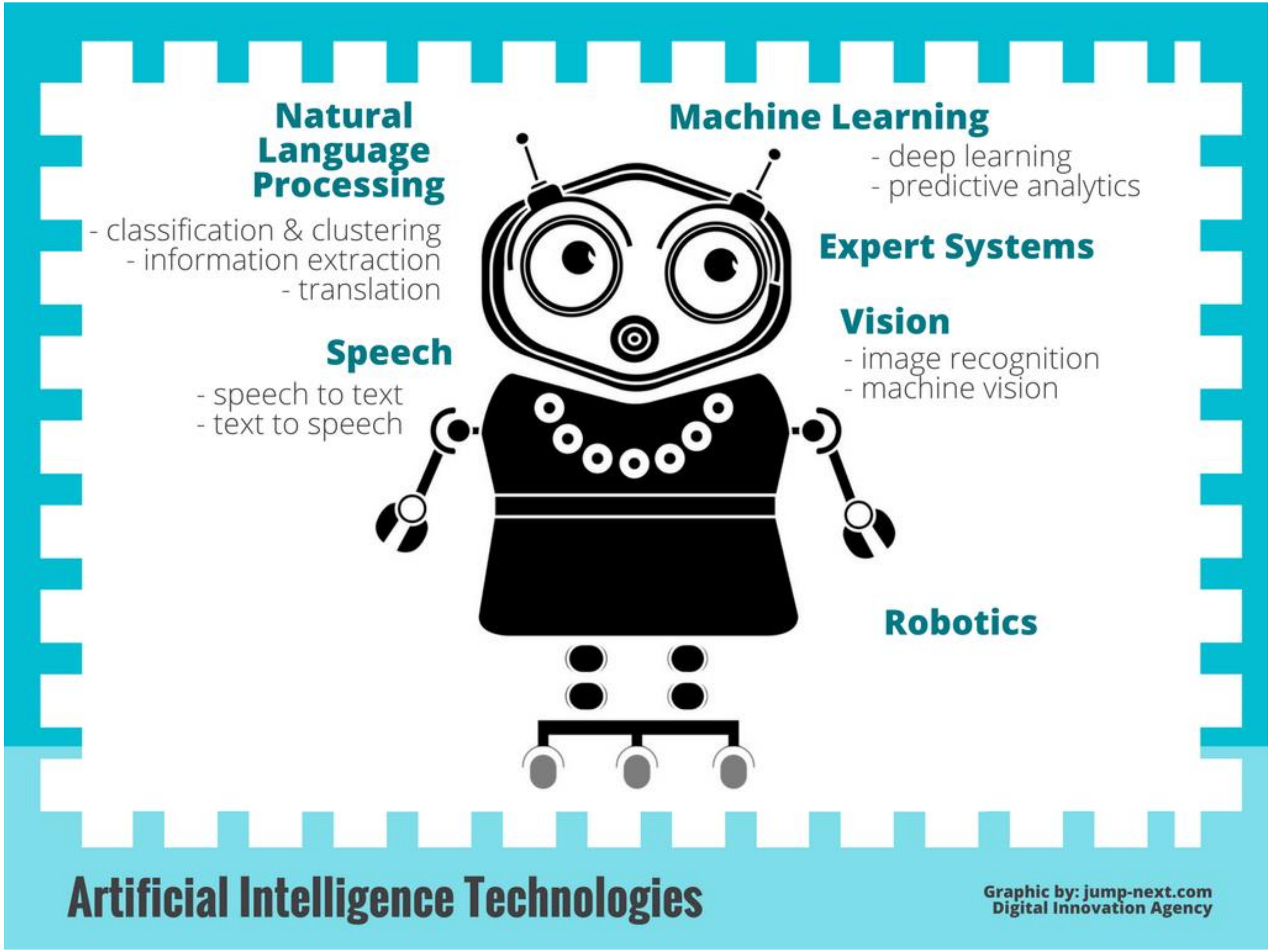
Speech Processing

Speech2Text & Text2Speech

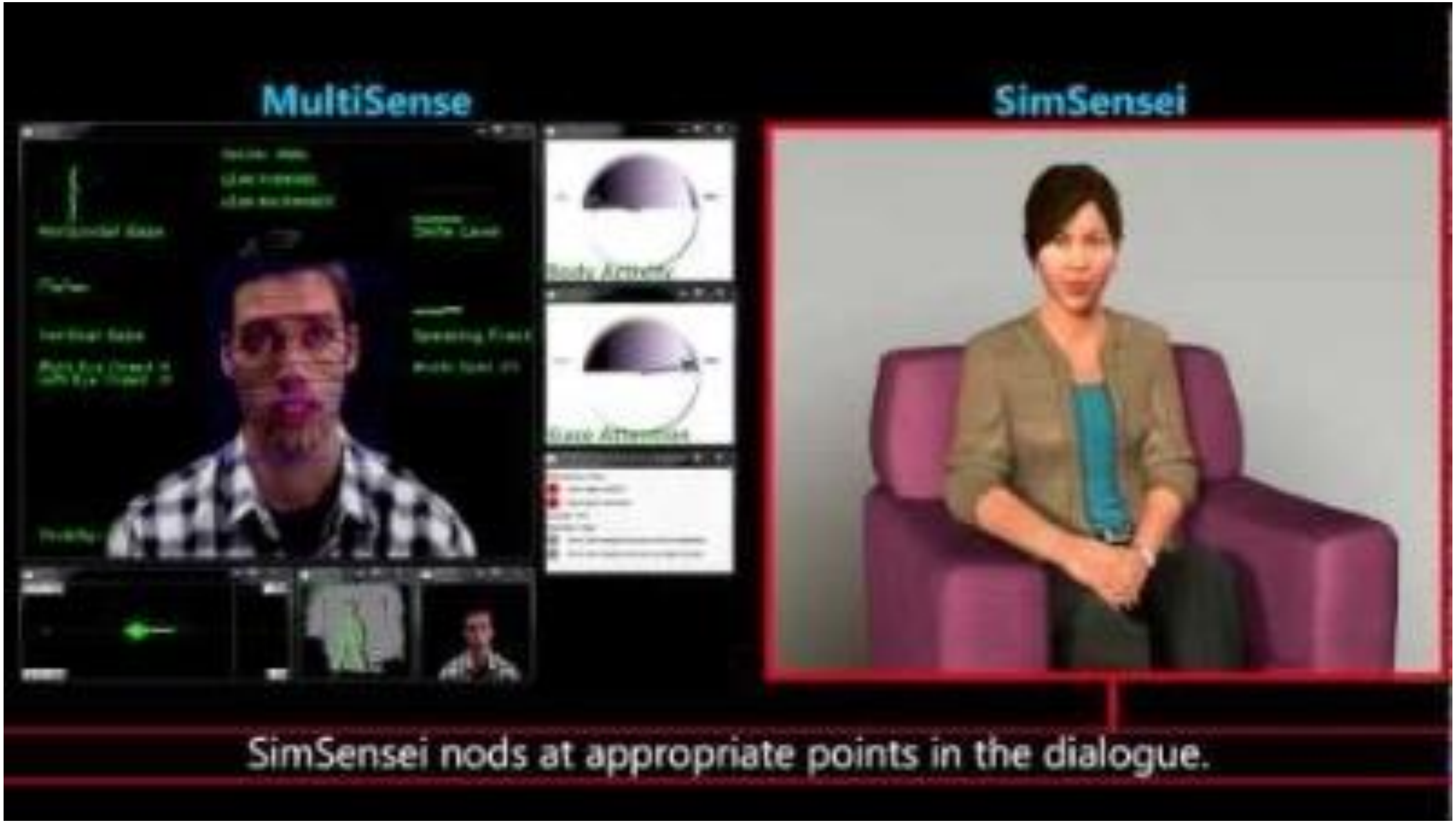
Language Processing

Making (voice) chatbots

Case Study 1: Robot

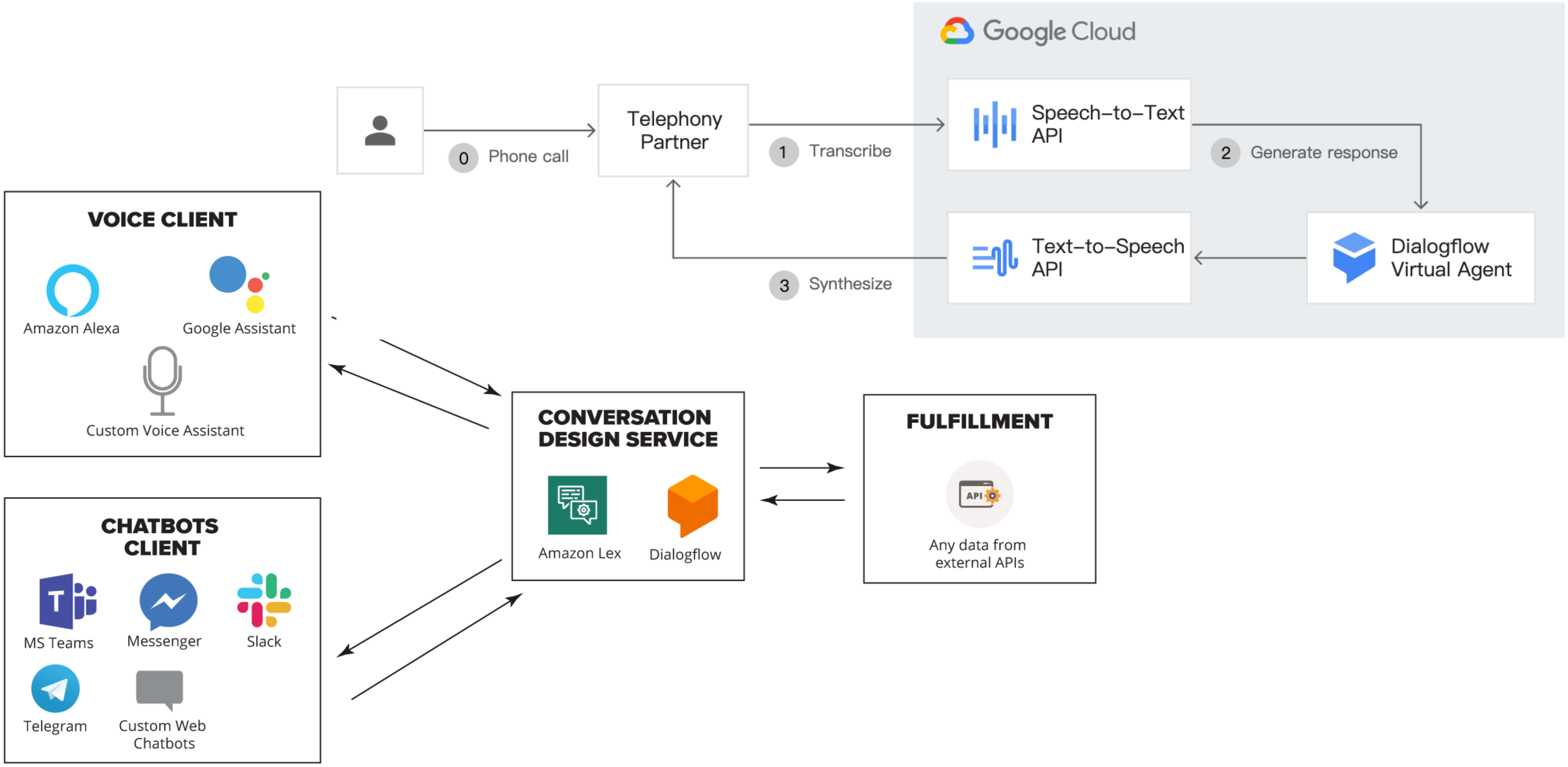


Case Study 2: SimSensi



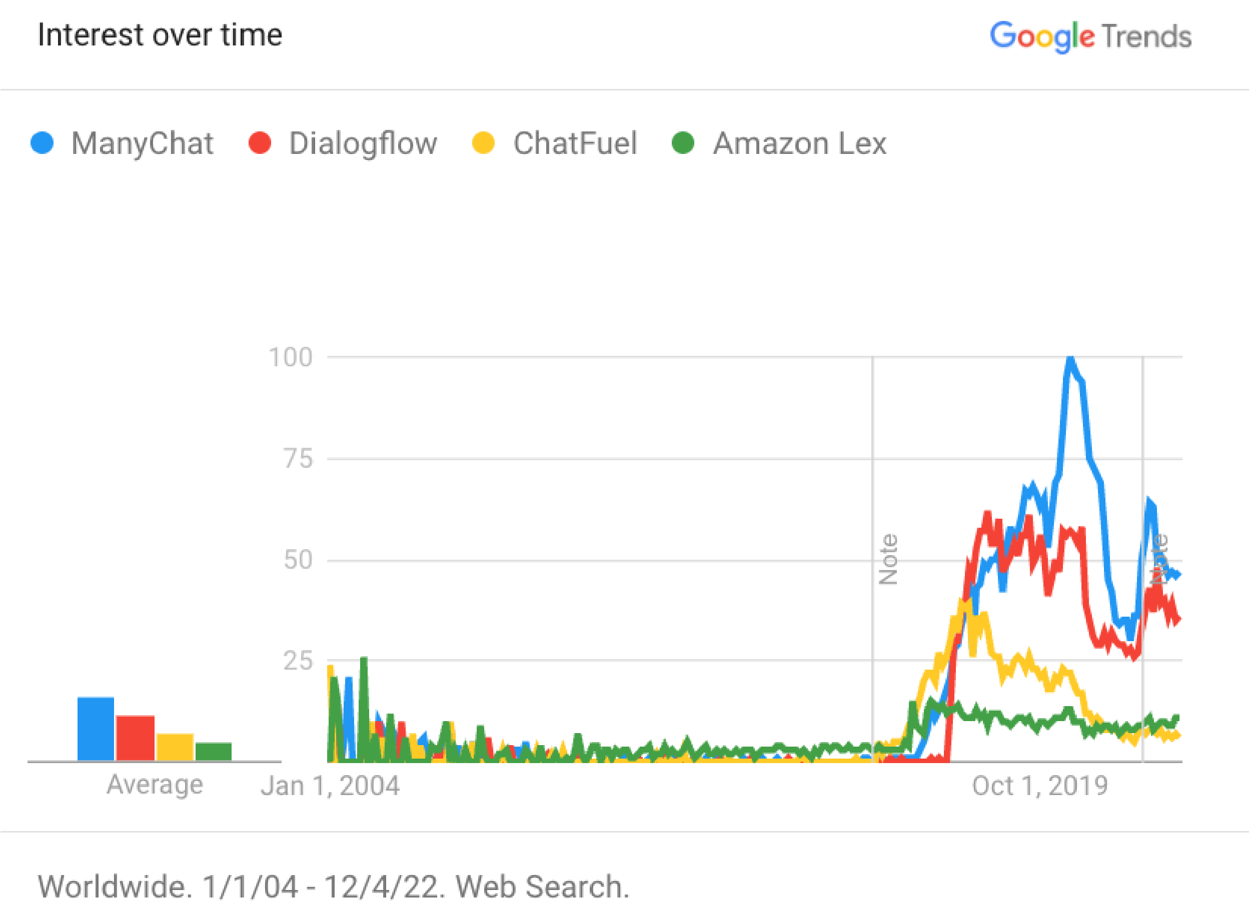
https://www.youtube.com/watch?v=ejczMs6b1Q4

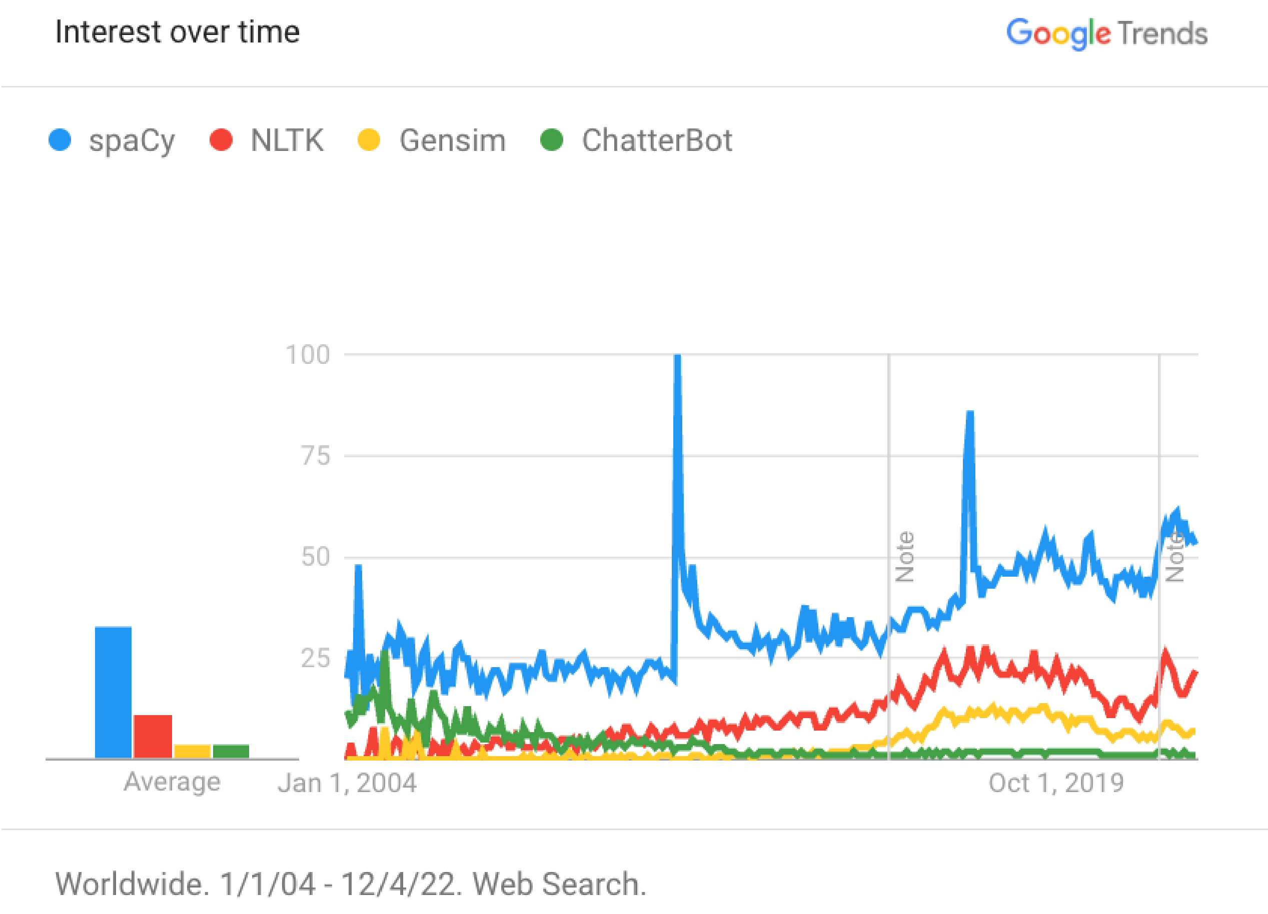
A voicebot is actually just a chatbot



Chatbot Services/Engines

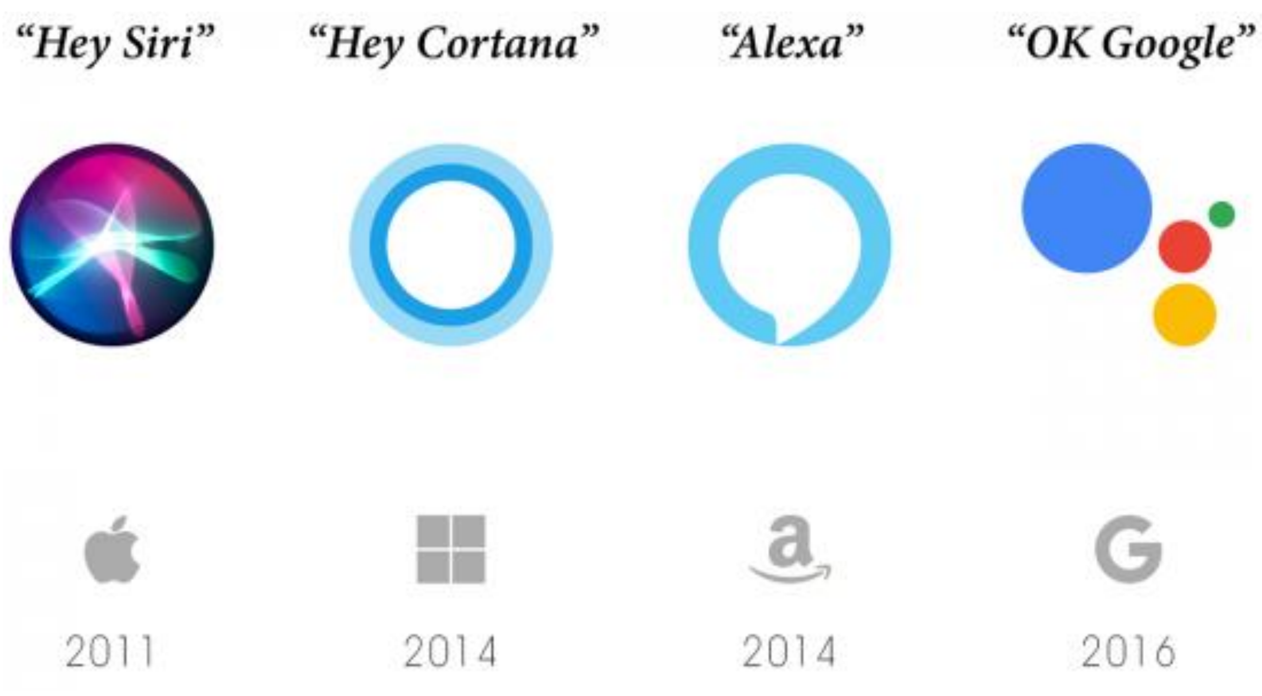
You can design your Q&A or outsource to other chatbots(!)





State-of-the-art Chatbots

Task-oriented:

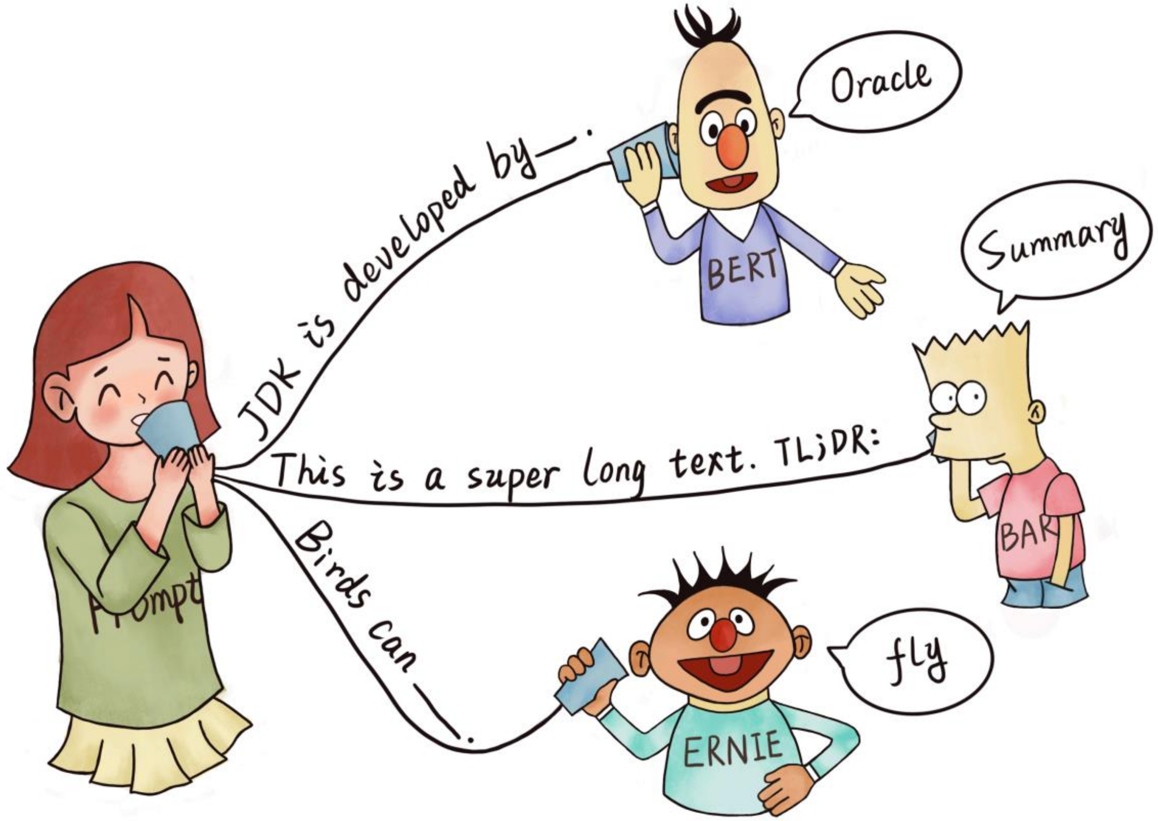


Non-task-oriented:

Prompt Programming

This is the best we can do w/o fine-tuning a LM on our data:

Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing



Q: Japan	This movie is the best I've ever watched.
A: Tokyo	On a scale from 0 to 100, I will rate it a 95!
Q: Frence	This movie is just ok.
A: Paris	On a scale from 0 to 100, I will rate it a 50.
Q: Taiwan	This movie sucks.
A: Taipei	On a scale from 0 to 100, I will rate it a 0.

Topics for today

Audio Processing

Extraction of sound features

Speech Processing

Speech2Text & Text2Speech

Language Processing

Making (voice) chatbots

Game Over