Introduction to Automatic Speech and Speaker Recognition

Sadaoki Furui
Tokyo Institute of Technology
Department of Computer Science
furui@cs.titech.ac.jp

Major speech recognition applications

- **Conversational systems** for accessing information services

  (e.g. automatic flight status or stock quote information systems)

- Systems for **transcribing, understanding and summarizing** ubiquitous speech documents

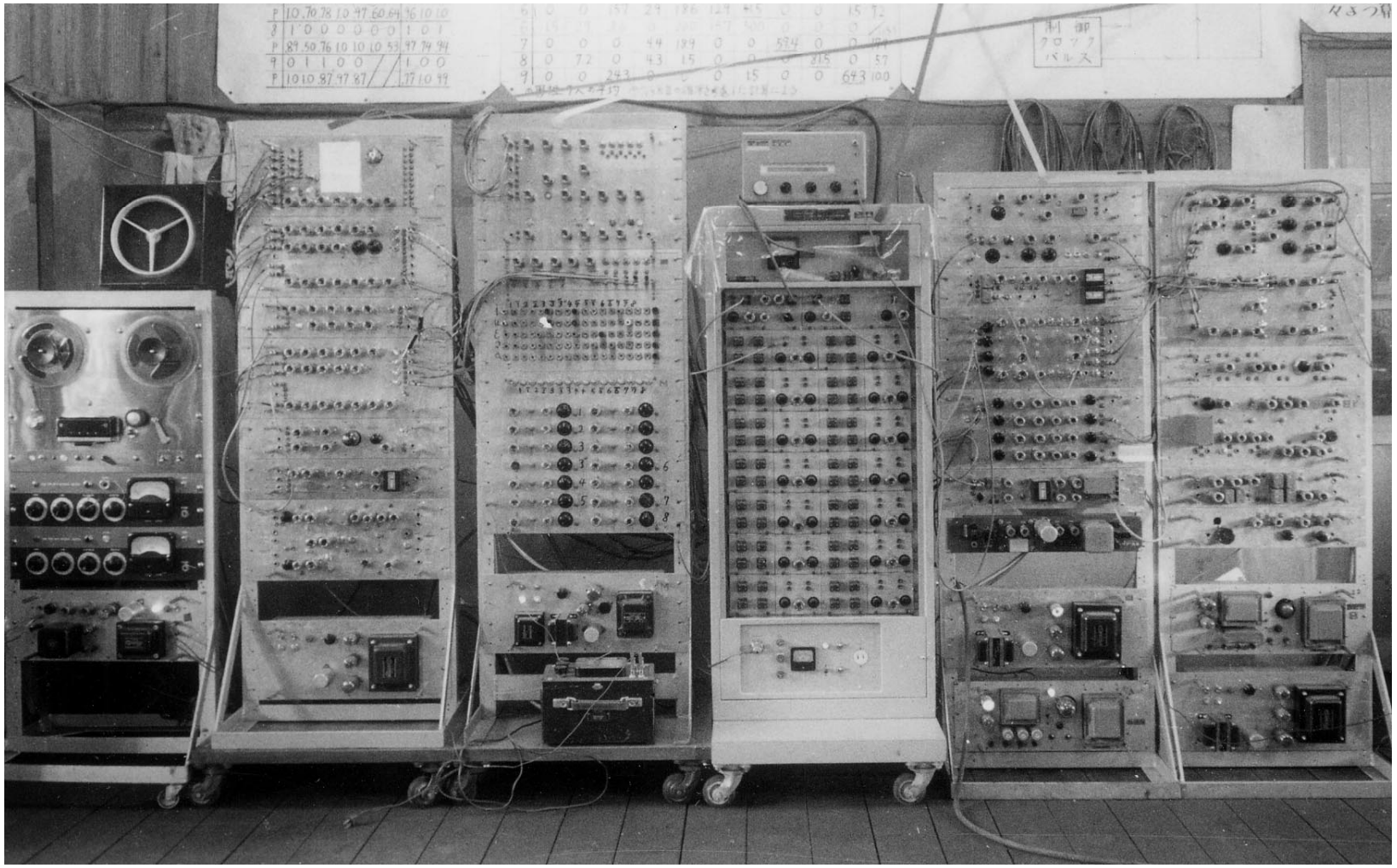
  (e.g. broadcast news, meetings, lectures, presentations, congressional records, court records, and voicemails)

Radio Rex – 1920's ASR

A sound-activated toy dog named "Rex" (from Elmwood Button Co.) could be called by name from his doghouse by name.

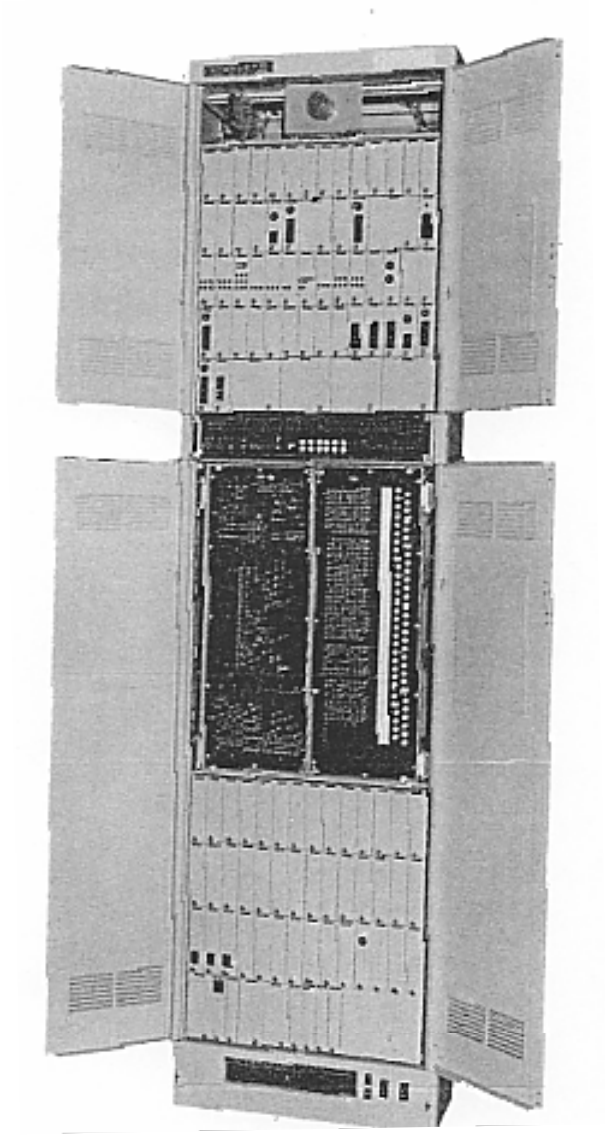
Front view of the spoken digit recognizer

(J. Suzuki & K. Nakata, Radio Research Labs, Japan, 1961)



Photograph of the Japanese spoken digit recognizer

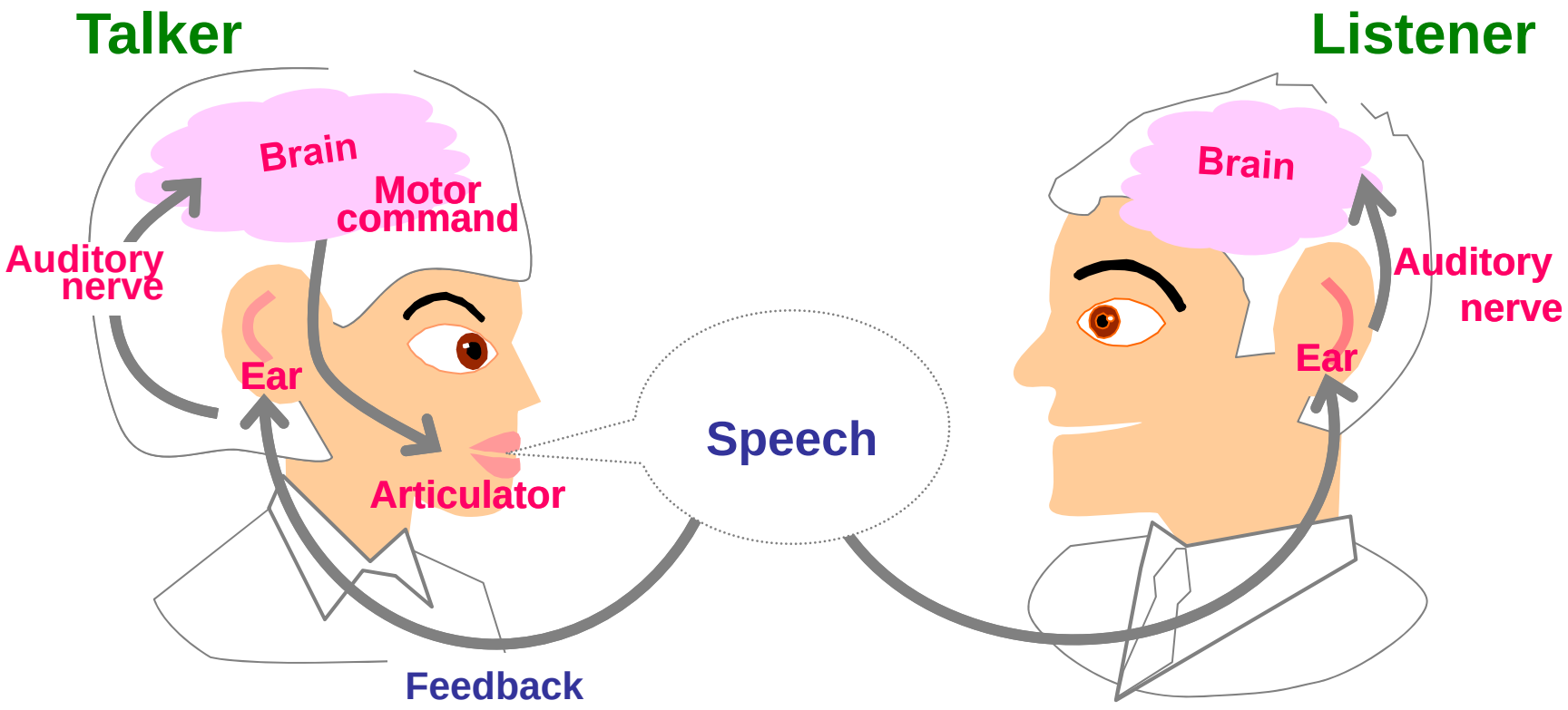
(K. Nagata, Y. Kato and S. Chiba, NEC Labs, Japan, 1963)



“Julie” doll with speech synthesis and recognition technology, produced by Worlds of Wonder in conjunction with Texas Instruments (1987)

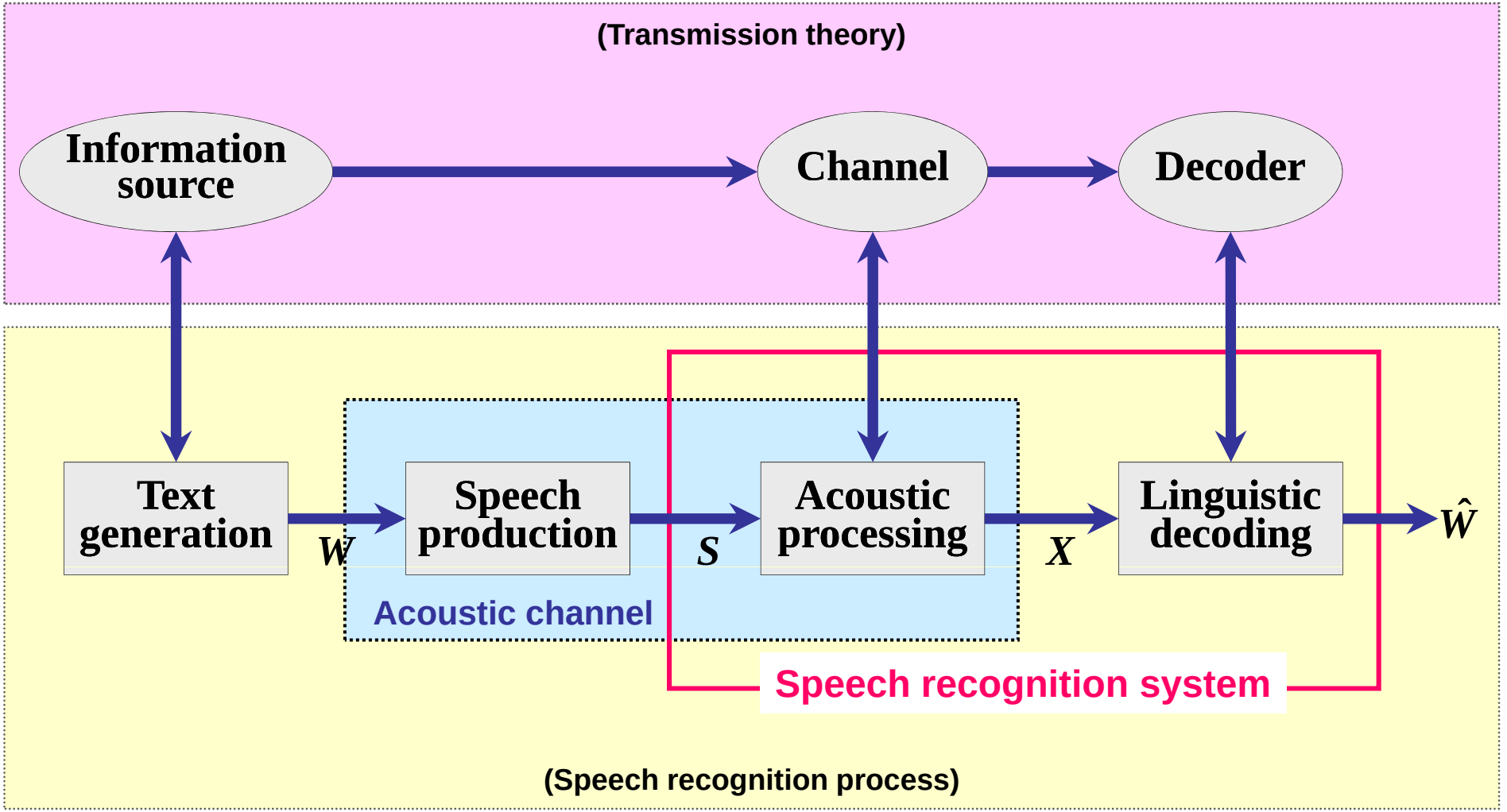
Now

Speech chain



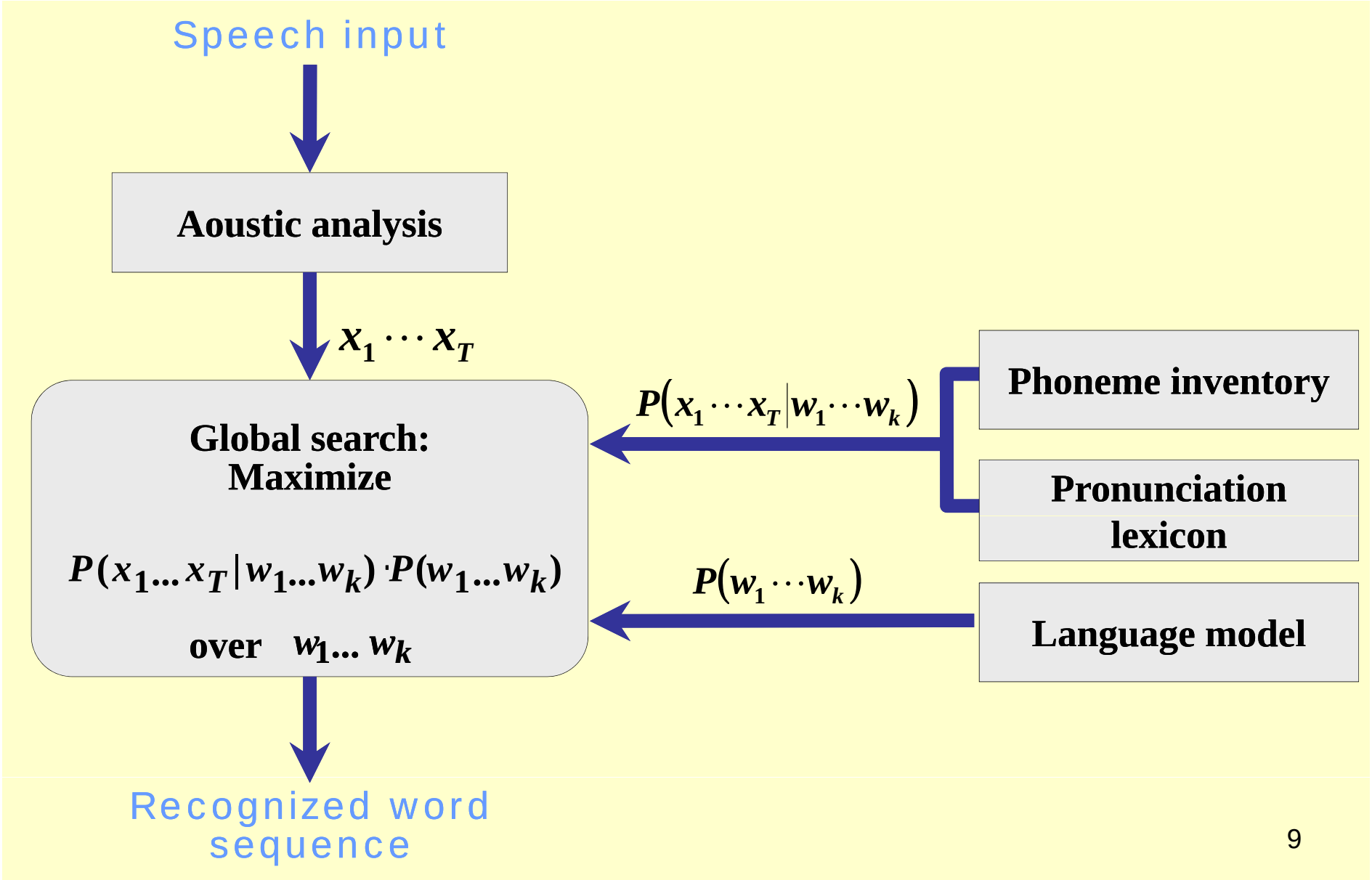
Linguistic process	Physiological process	Physical (acoustic) process	Physiological process	Linguistic process
Discrete	Continuous			Discrete

Structure of speech production and recognition system based on information transmission theory

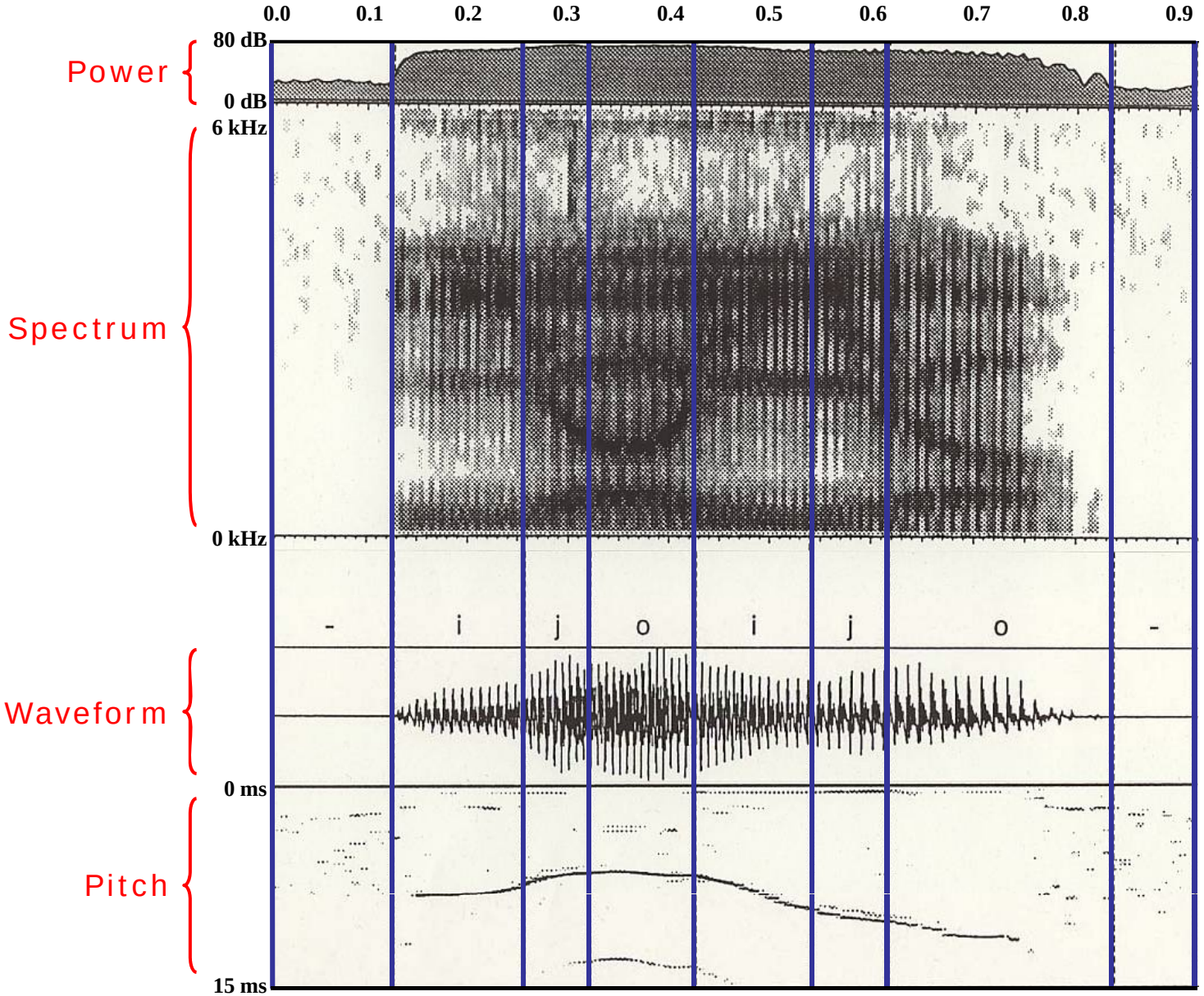


$$\hat{W} = \arg\max_{W} P(W|X) = \arg\max_{W} \frac{P(X|W)P(W)}{P(X)}$$

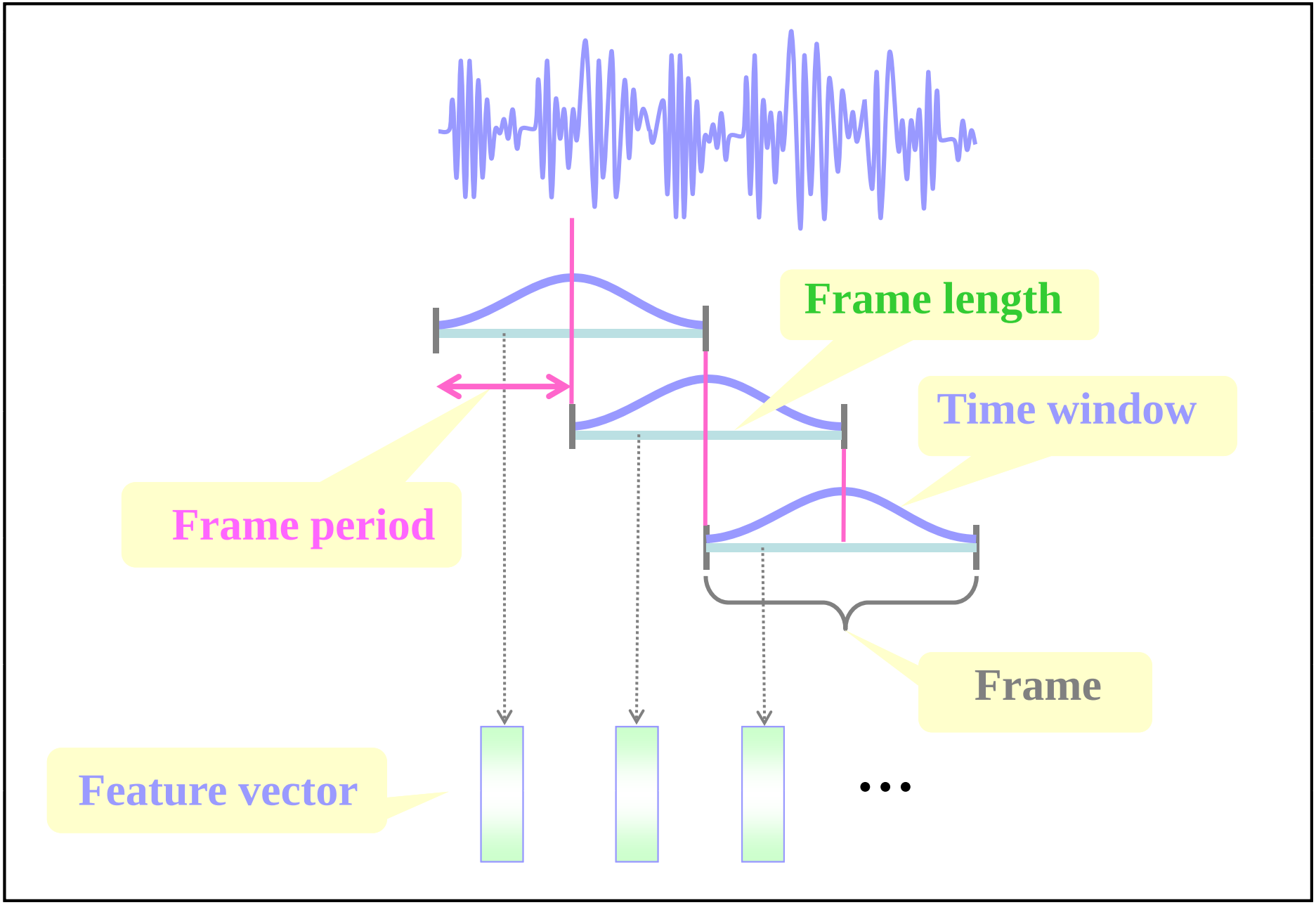
Mechanism of state-of-the-art speech recognizers



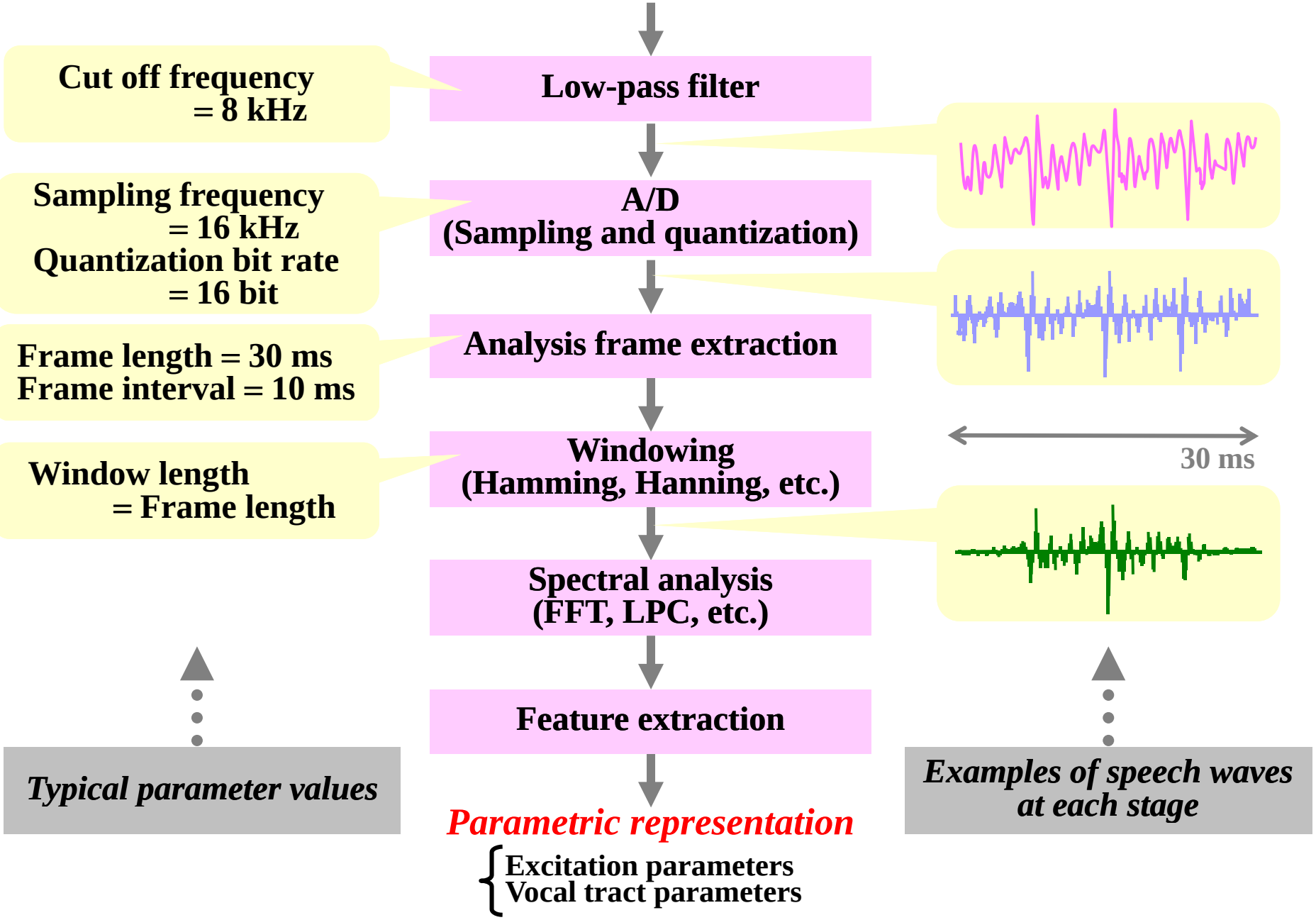
Speech energy, sound spectrogram, waveform and pitch



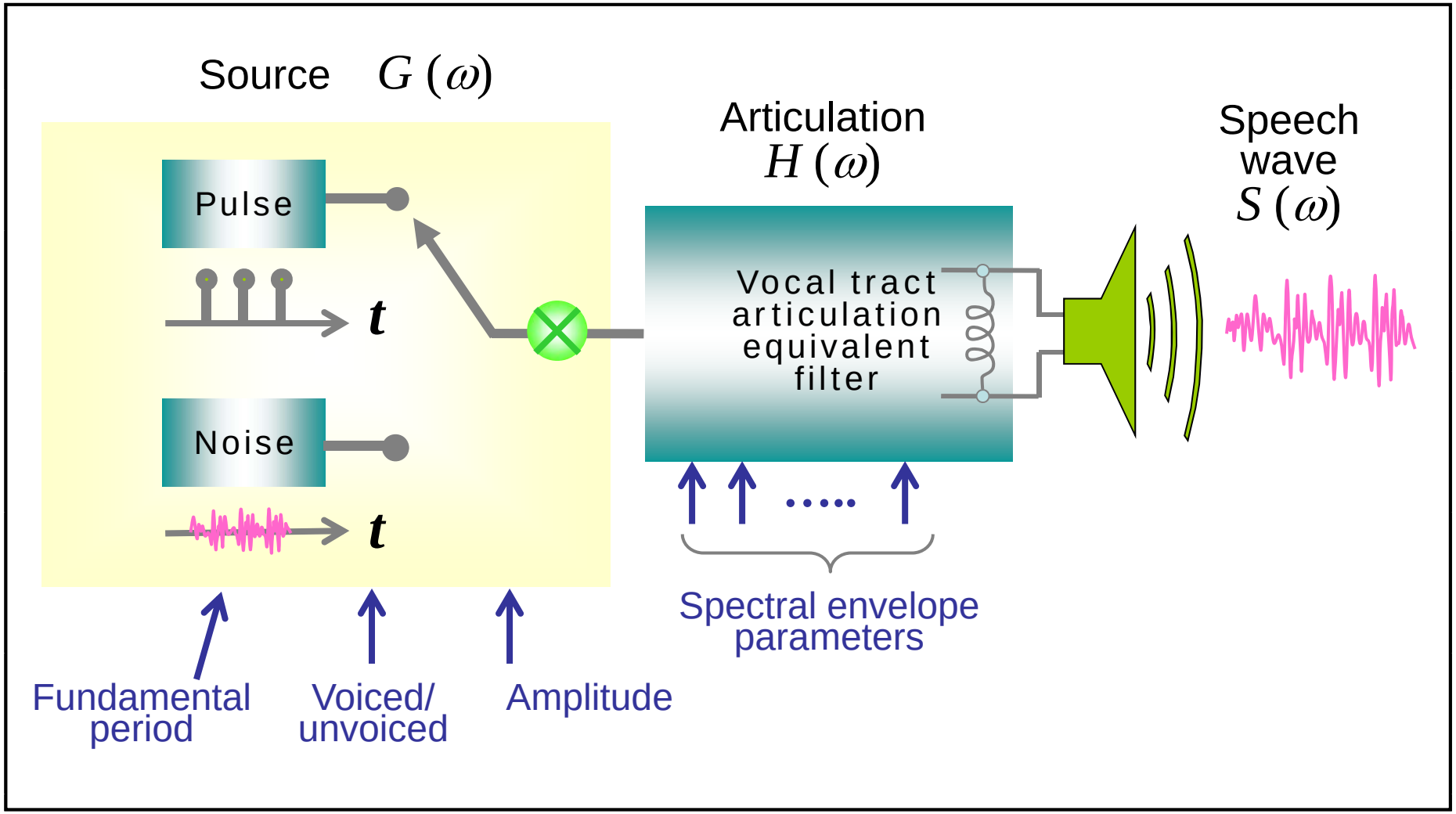
Feature vector (short-time spectrum) extraction from speech



Block diagram of a typical speech analysis procedure

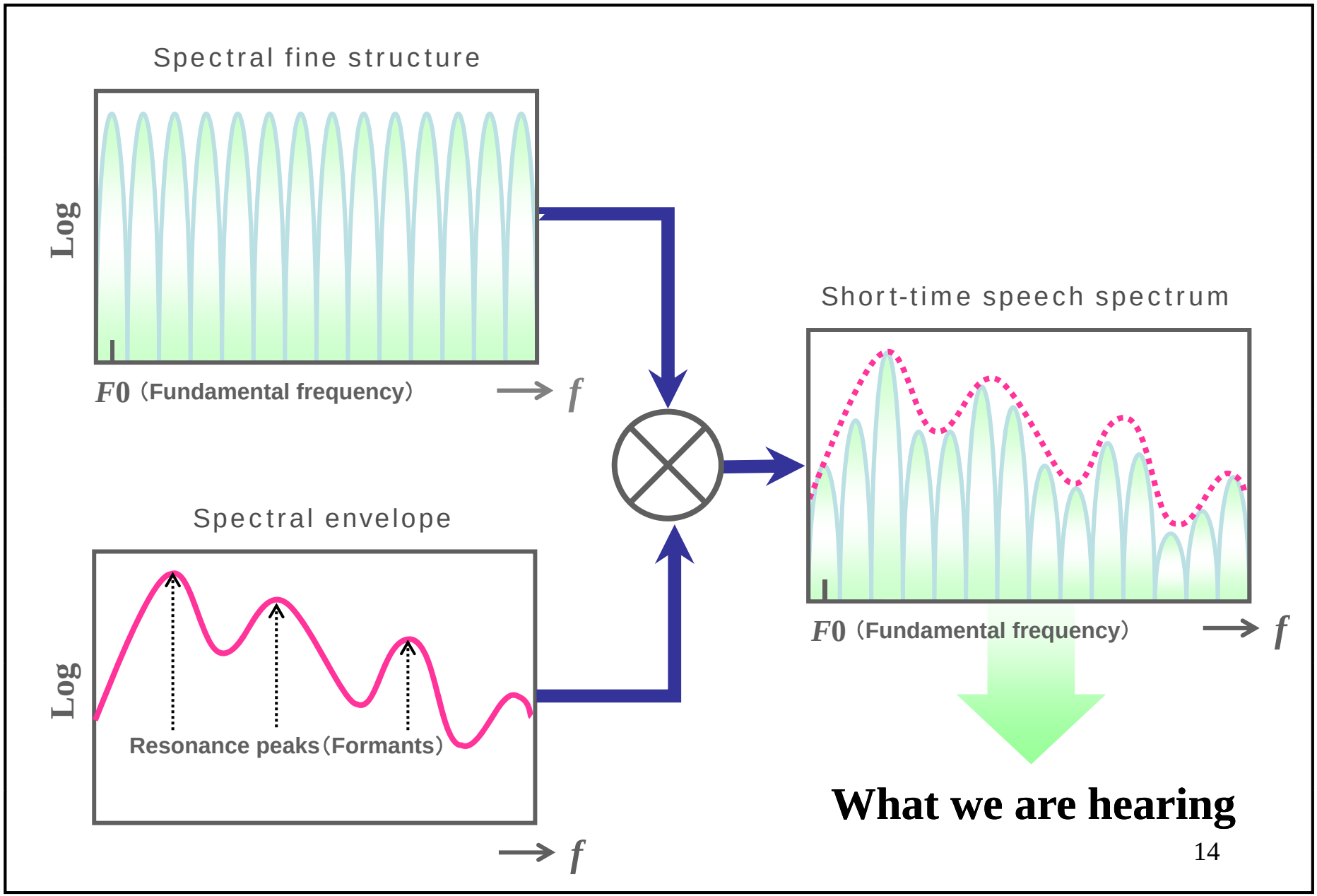


Linear separable equivalent circuit model of the speech production mechanism

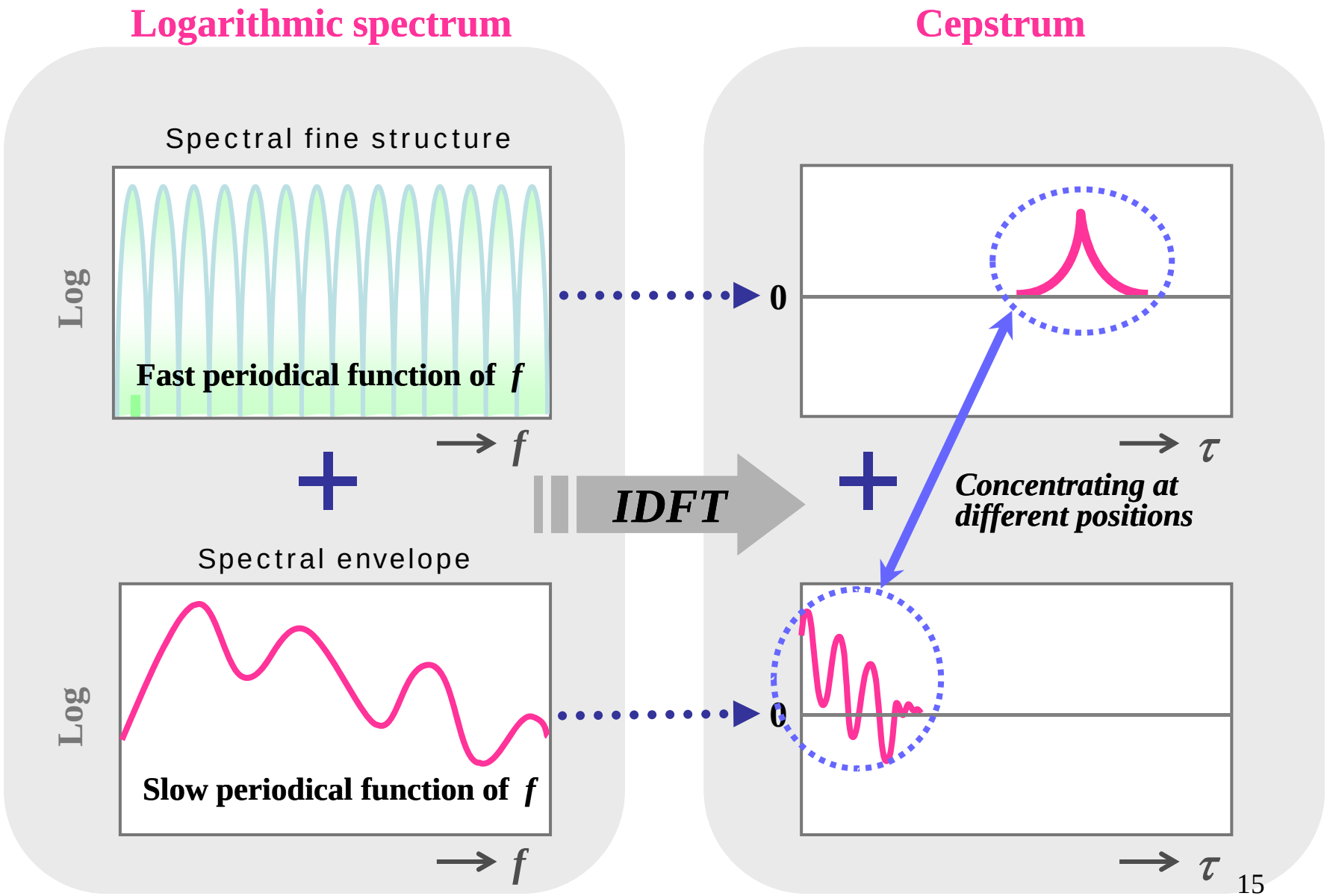


$$S(\omega) = G(\omega) \cdot H(\omega)$$

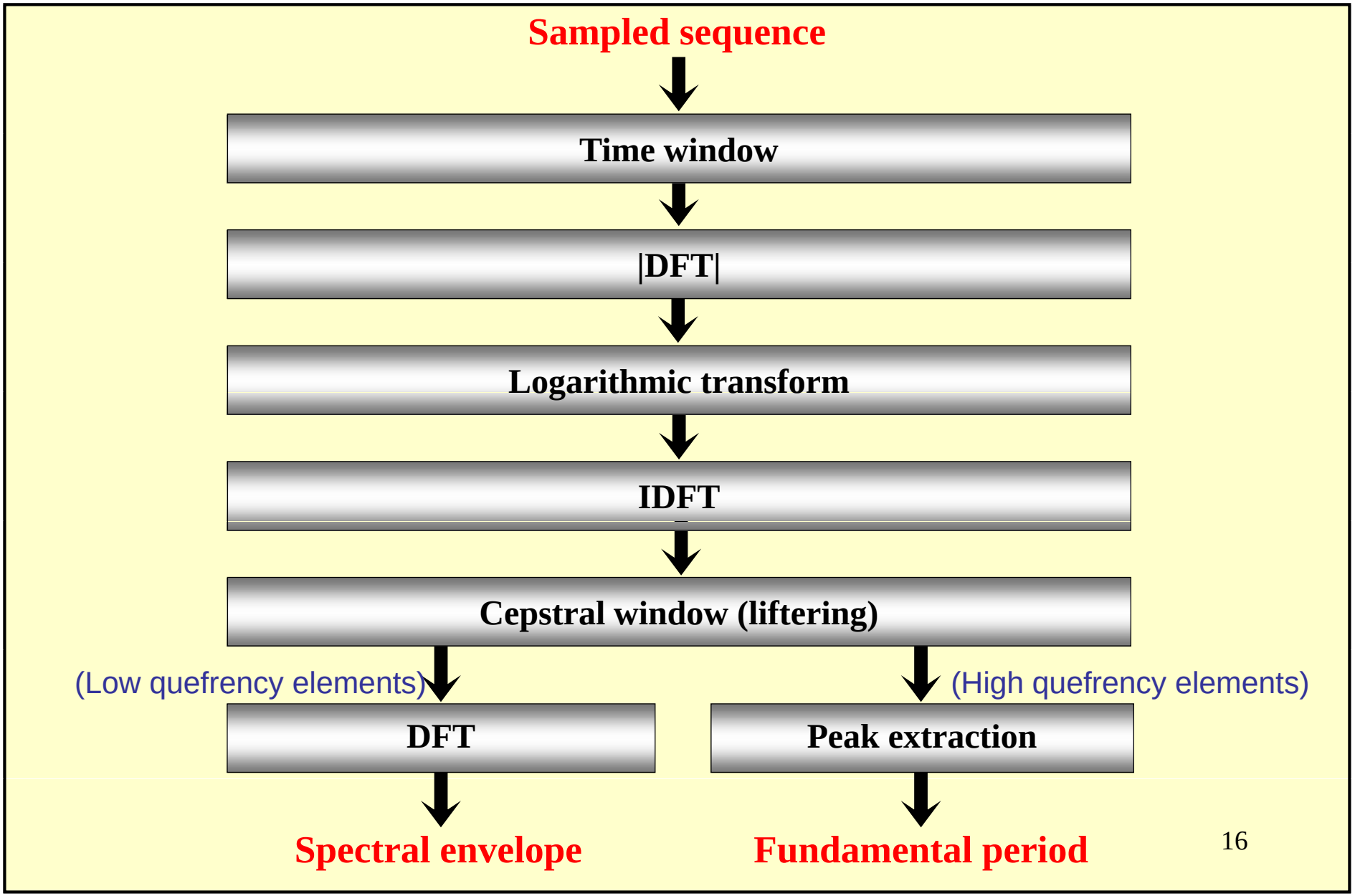
Spectral structure of speech



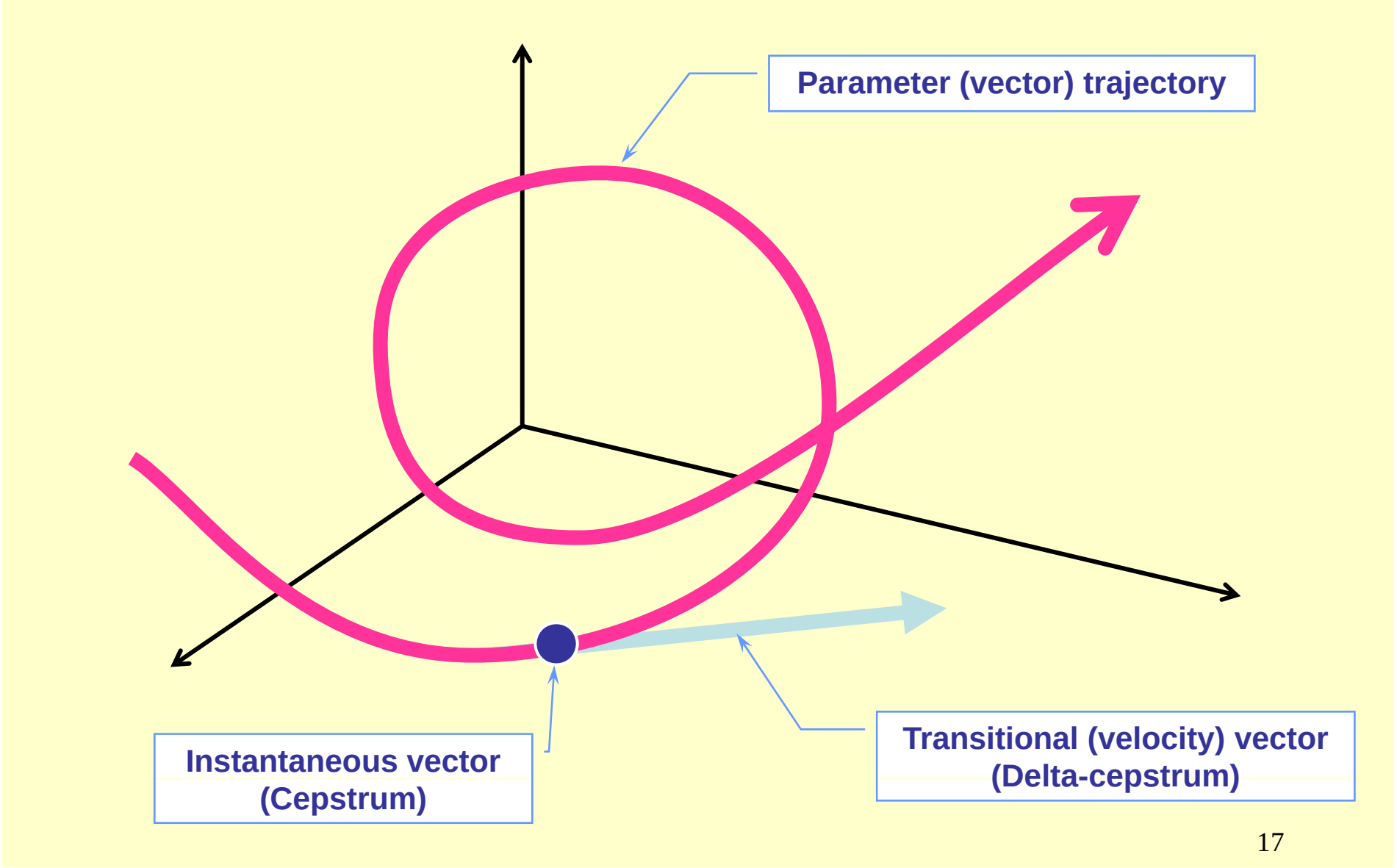
Relationship between logarithmic spectrum and cepstrum



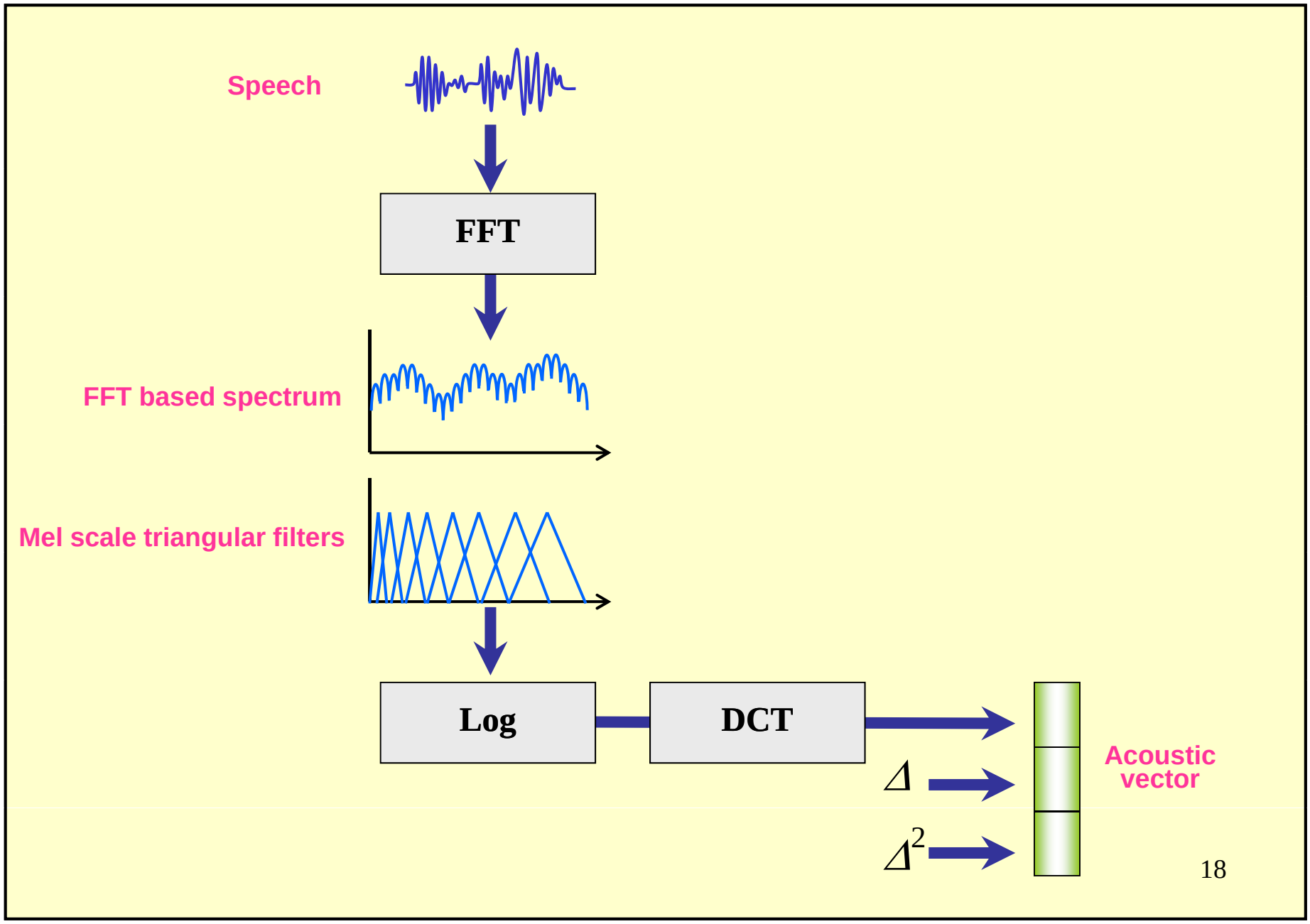
Block diagram of cepstrum analysis for extracting spectral envelope and fundamental period



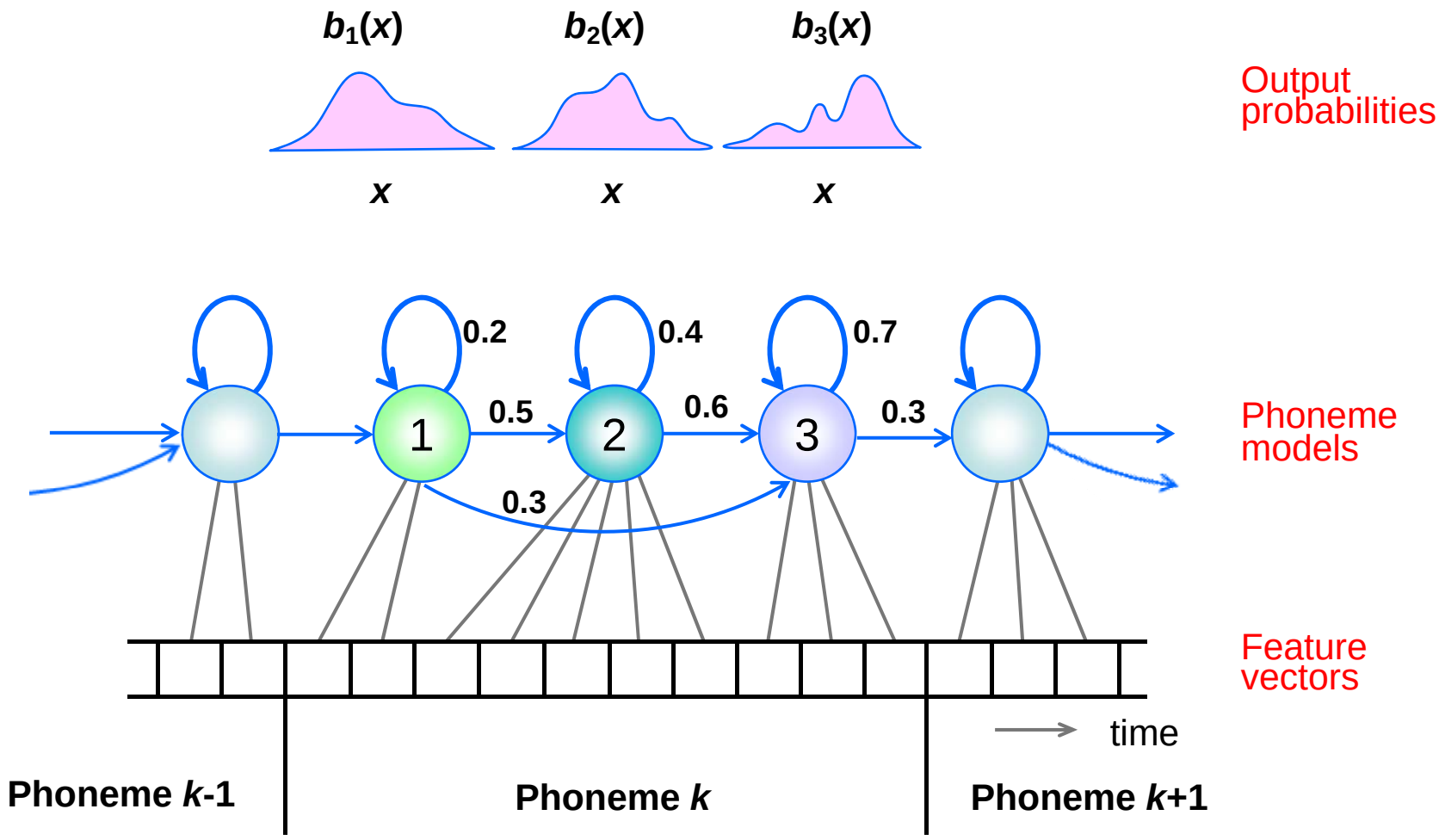
Cepstrum and delta-cepstrum coefficients



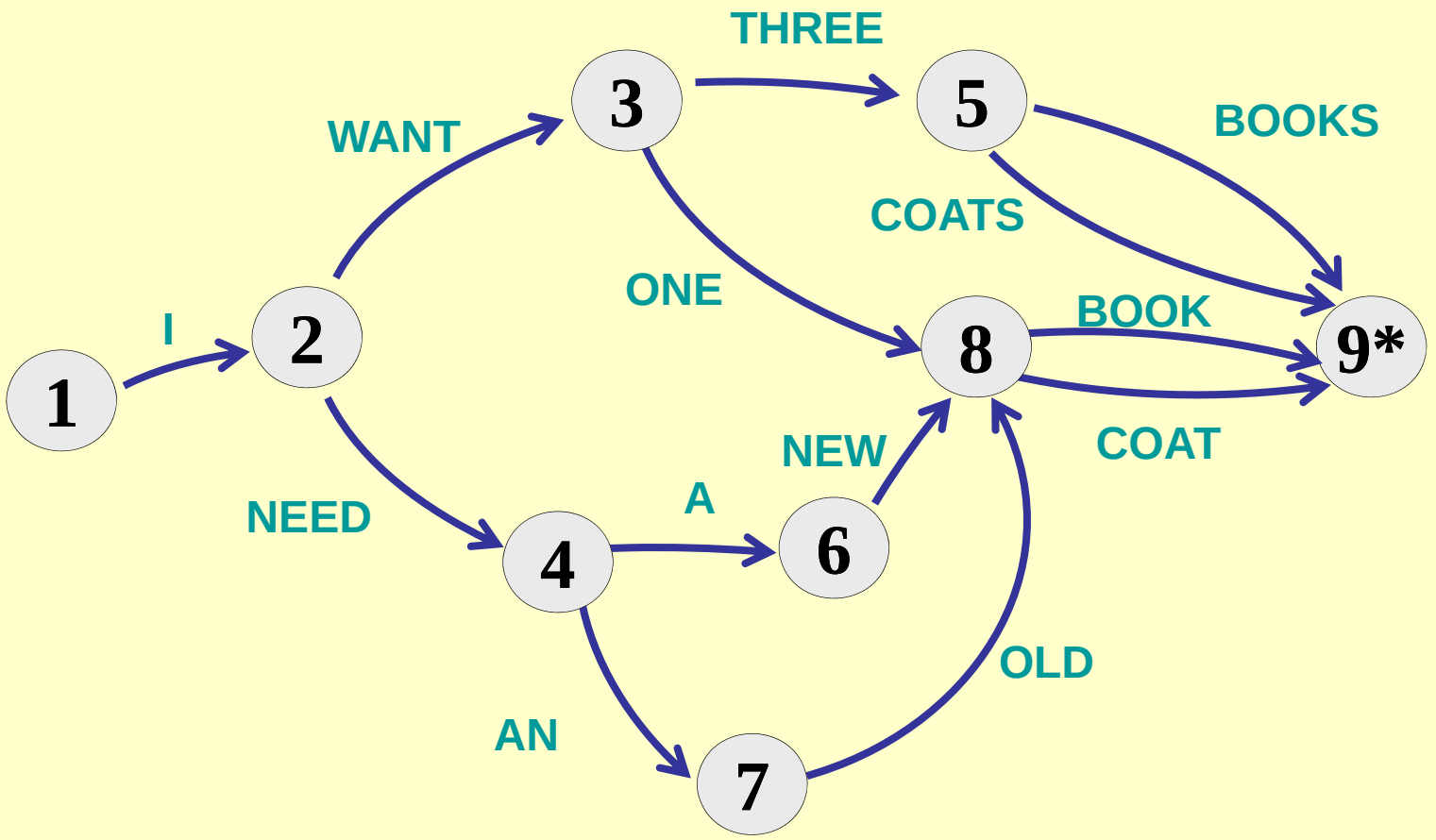
MFCC-based front-end processor



Structure of phoneme HMMs



An example of FSN (Finite State Network) grammar



1.	I	5.	ONE	9.	BOOKS	13.	OLD
2.	WANT	6.	A	10.	COAT		
3.	NEED	7.	AN	11.	COATS		
4.	THREE	8.	BOOK	12.	NEW		

Statistical language modeling

Probability of the word sequence $w_1^k = w_1 w_2 \ldots w_k$:

$$P(w_1^k) = \prod_{i=1}^{k} P(w_i \mid w_1 w_2 \ldots w_{i-1}) = \prod_{i=1}^{k} P(w_i \mid w_1^{i-1})$$

$$P(w_i \mid w_1^{i-1}) = N(w_1^i) \,/\, N(w_1^{i-1})$$

where $N(w_1^i)$ is the number of occurrences of the string w_1^i in the given training corpus.

Approximation by Markov processes:

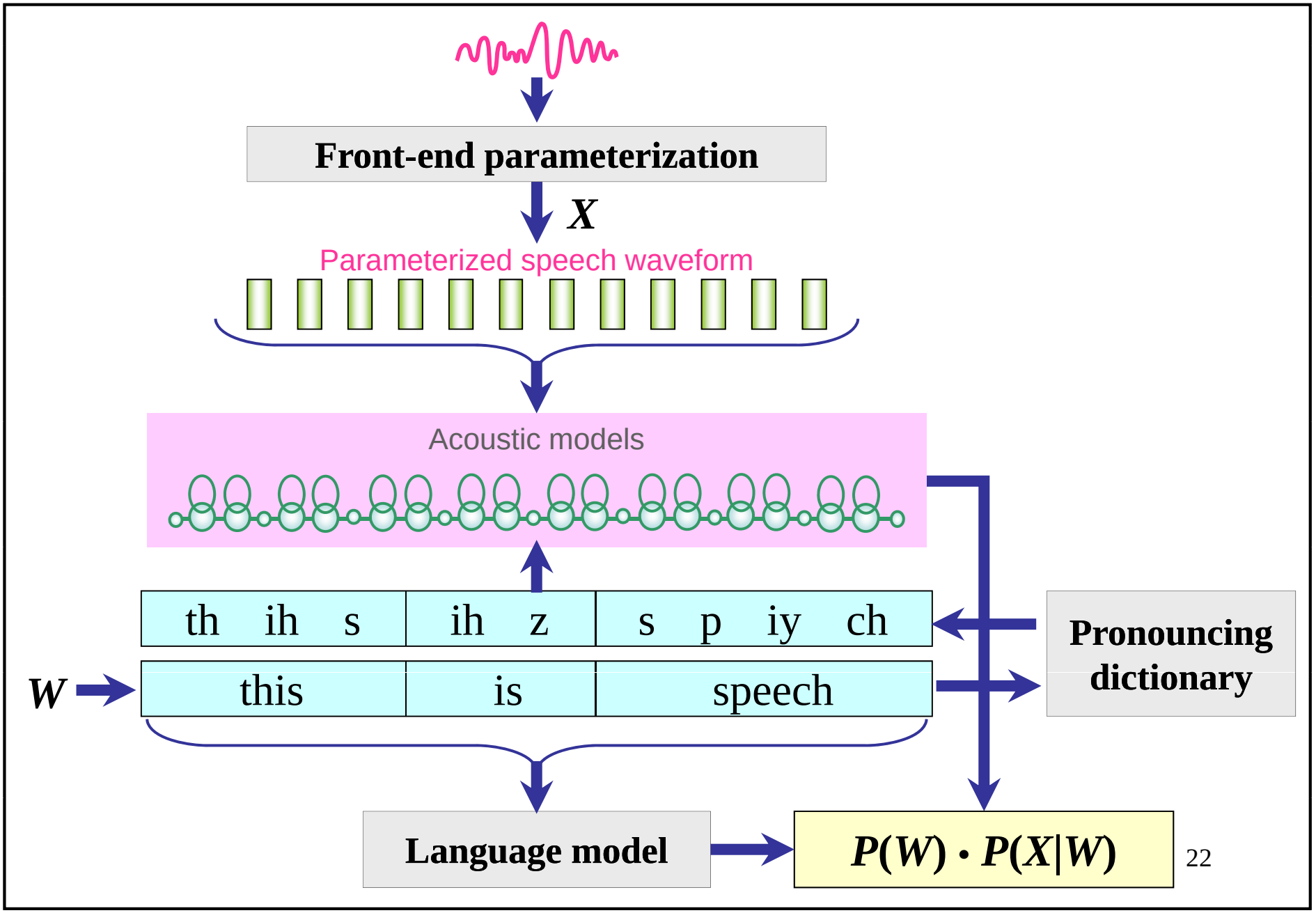
Bigram model $\quad P(w_i \mid w_1^{i-1}) = P(w_i \mid w_{i-1})$

Trigram model $\quad P(w_i \mid w_1^{i-1}) = P(w_i \mid w_{i-2} w_{i-1})$

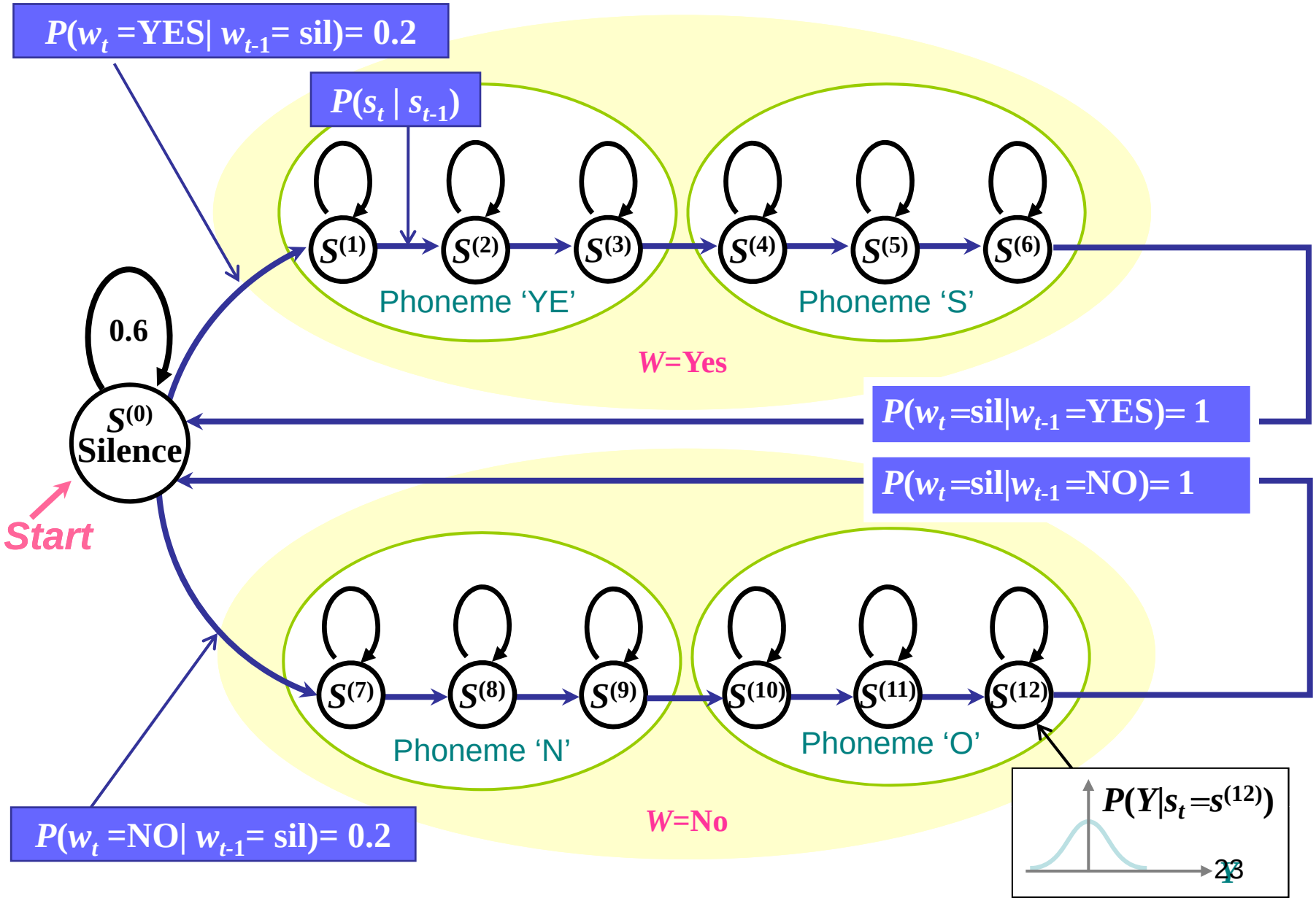
Smoothing of trigram by the deleted interpolation method:

$$P(w_i \mid w_{i-2} w_{i-1}) = \lambda_1 P(w_i \mid w_{i-2} w_{i-1}) + \lambda_2 P(w_i \mid w_{i-1}) + \lambda_3 P(w_i)$$

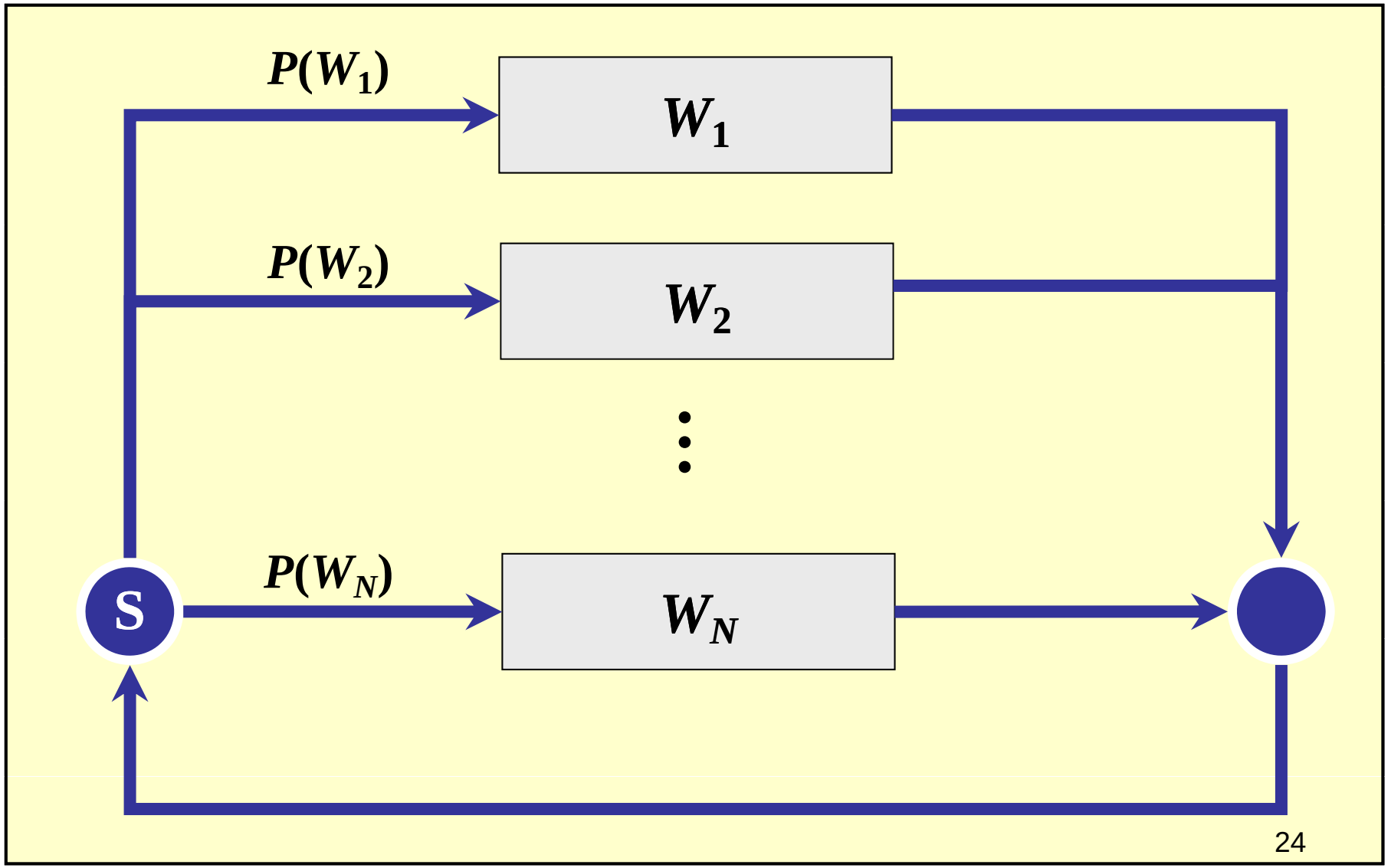
Overview of statistical speech recognition



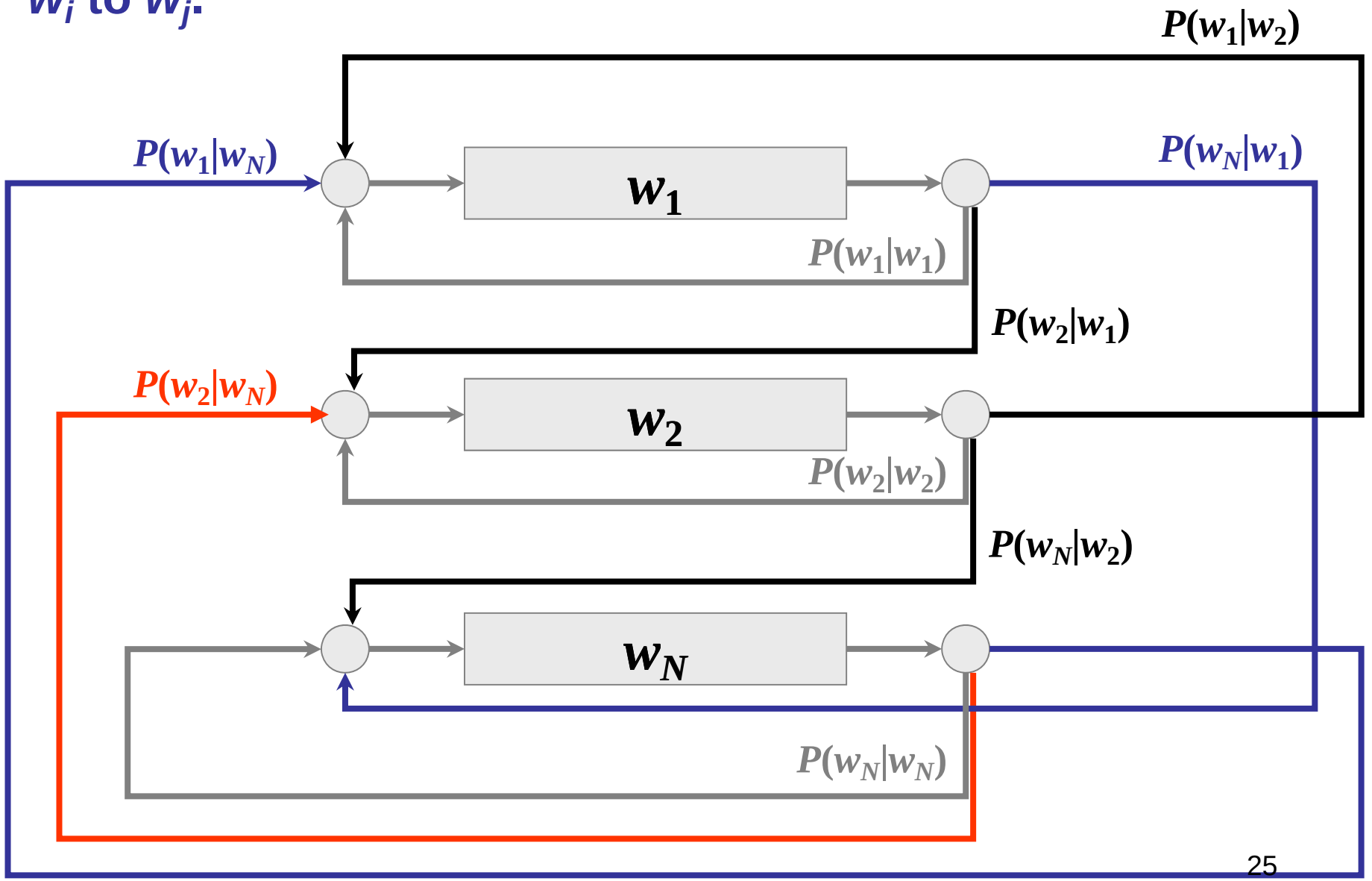
Complete Hidden Markov Model of a simple grammar



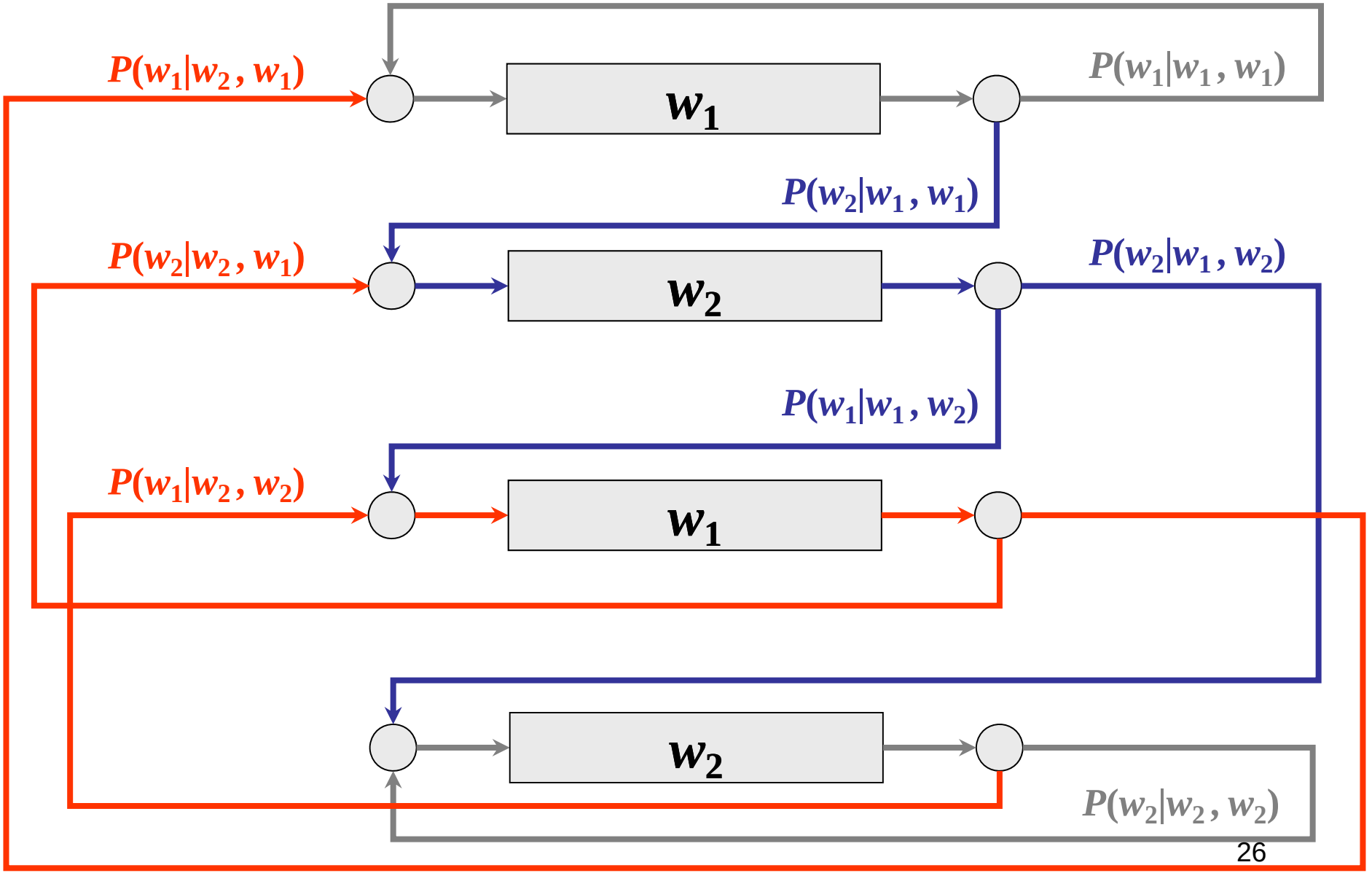
A unigram grammar network where the unigram probability is attached as the transition probability from starting state S to the first state of each word HMM.



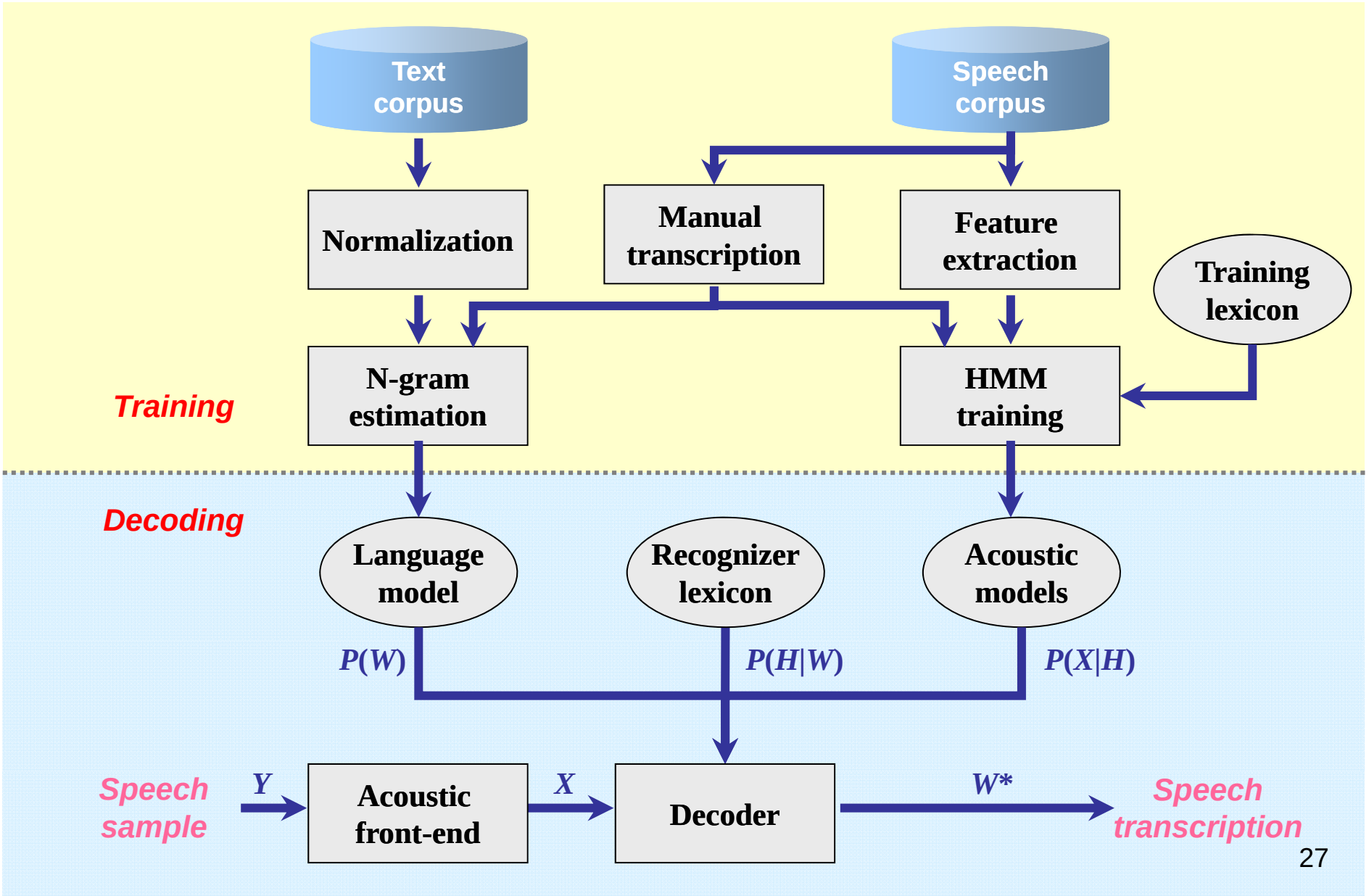
A bigram grammar network where the bigram probability $P(w_j|w_i)$ is attached as the transition probability from word w_i to w_j.



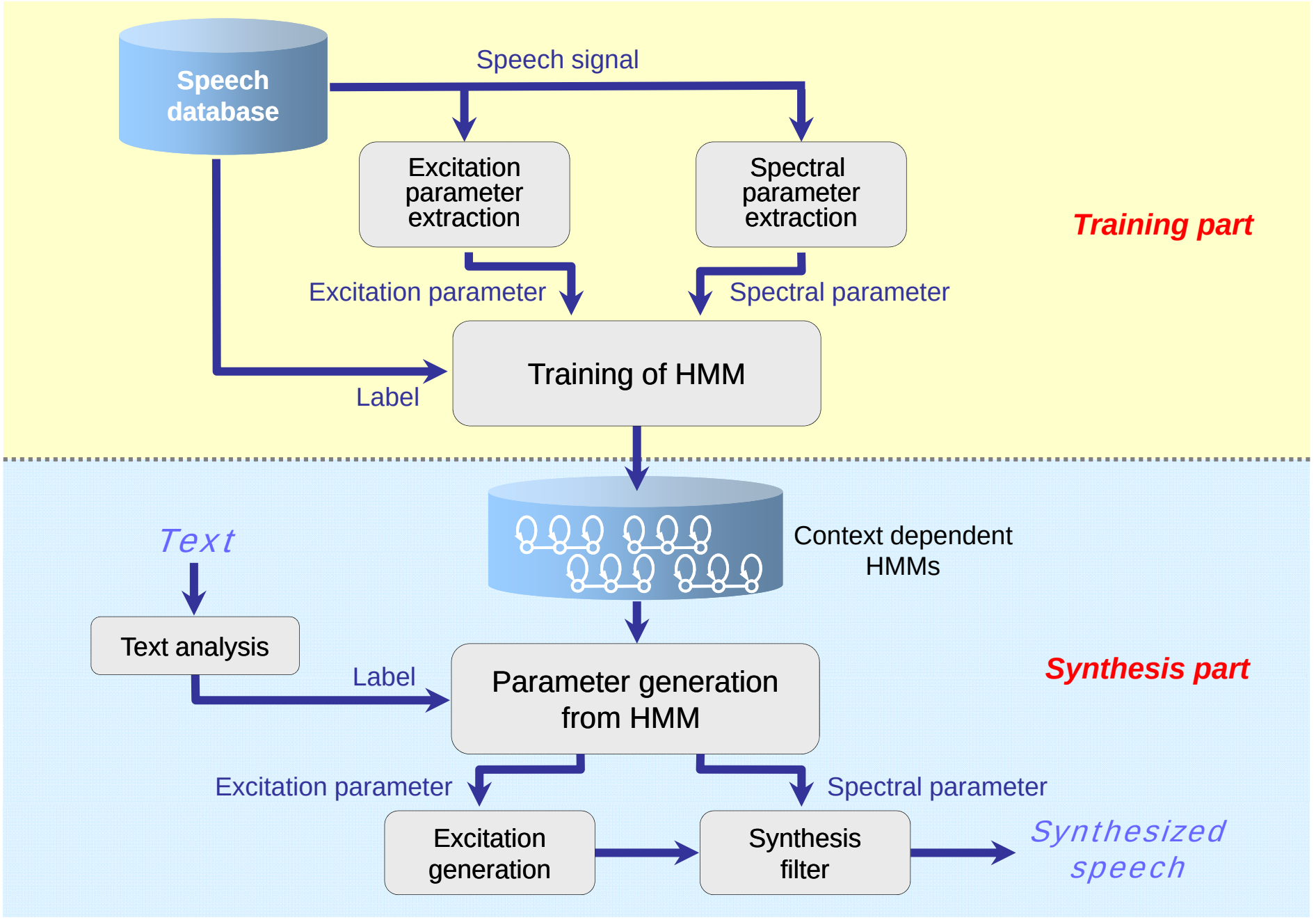
A trigram grammar network where the trigram probability $P(w_k|w_i, w_j)$ is attached to transition from grammar state w_i, w_j to the next word w_k. Illustrated here is a two-word vocabulary, so there are four grammar states in the trigram network.



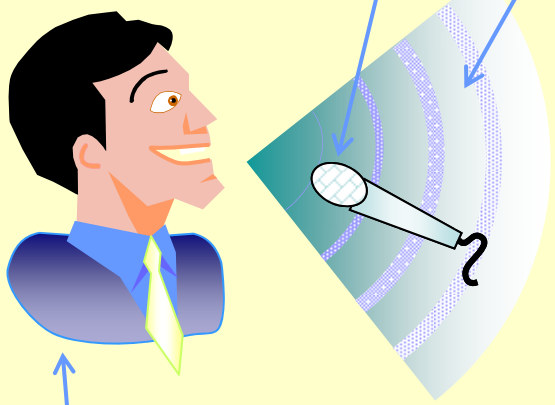
System diagram of a generic speech recognizer based on statistical models, including training and decoding processes and the main knowledge sources.



HMM-based speech synthesis system



Main causes of acoustic variation in speech



Microphone
- Distortion
- Electrical noise
- Directional characteristics

Noise
- Other speakers
- Background noise
- Reverberations

Distortion
Noise
Echoes
Dropouts

Channel

Speech recognition system

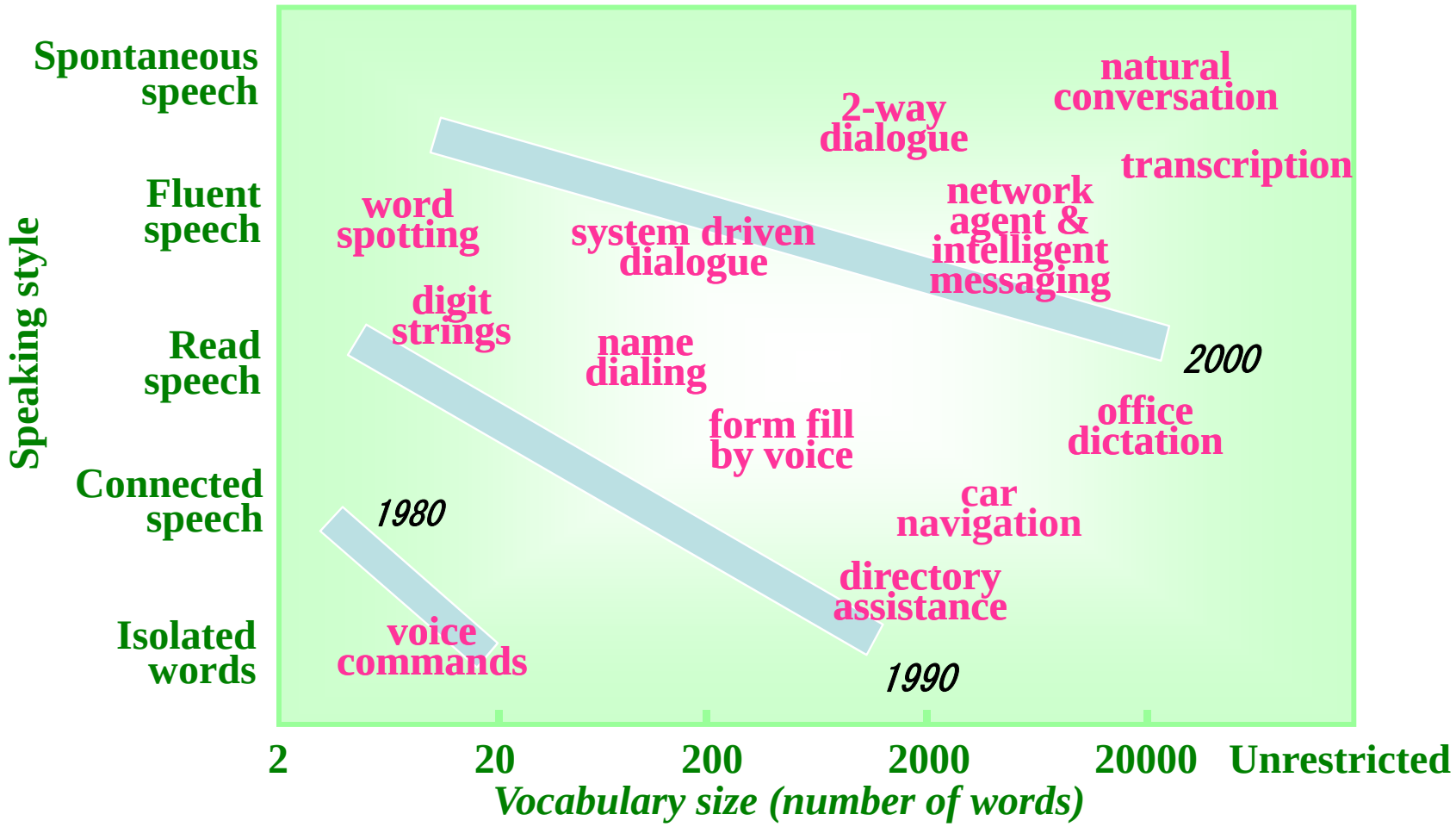
Speaker
- Voice quality
- Pitch
- Gender
- Dialect

Task/Context
- Man-machine dialogue
- Dictation
- Free conversation
- Interview

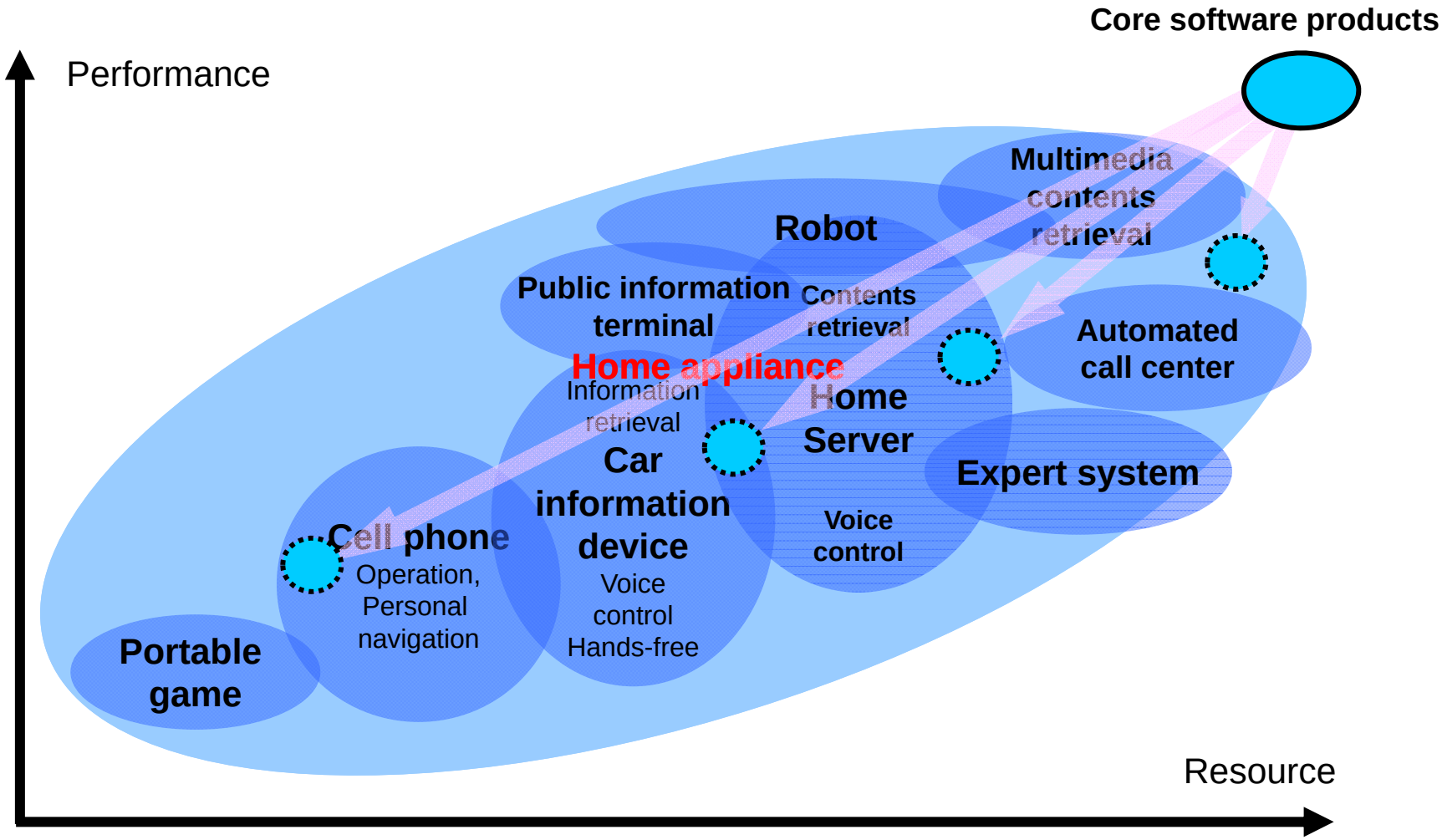
Speaking style
- Stress/Emotion
- Speaking rate
- Lombard effect

Phonetic/Prosodic context

Progress of speech recognition technology since 1980



Various speech applications



Speaker recognition

"voice key"

- **Speaker verification**: confirm the identity claim (banking transactions, database access services, security control for confidential information)
- **Speaker identification**: determine from registered speakers (criminal investigations)

- **Text-dependent** methods
- **Text-independent** methods

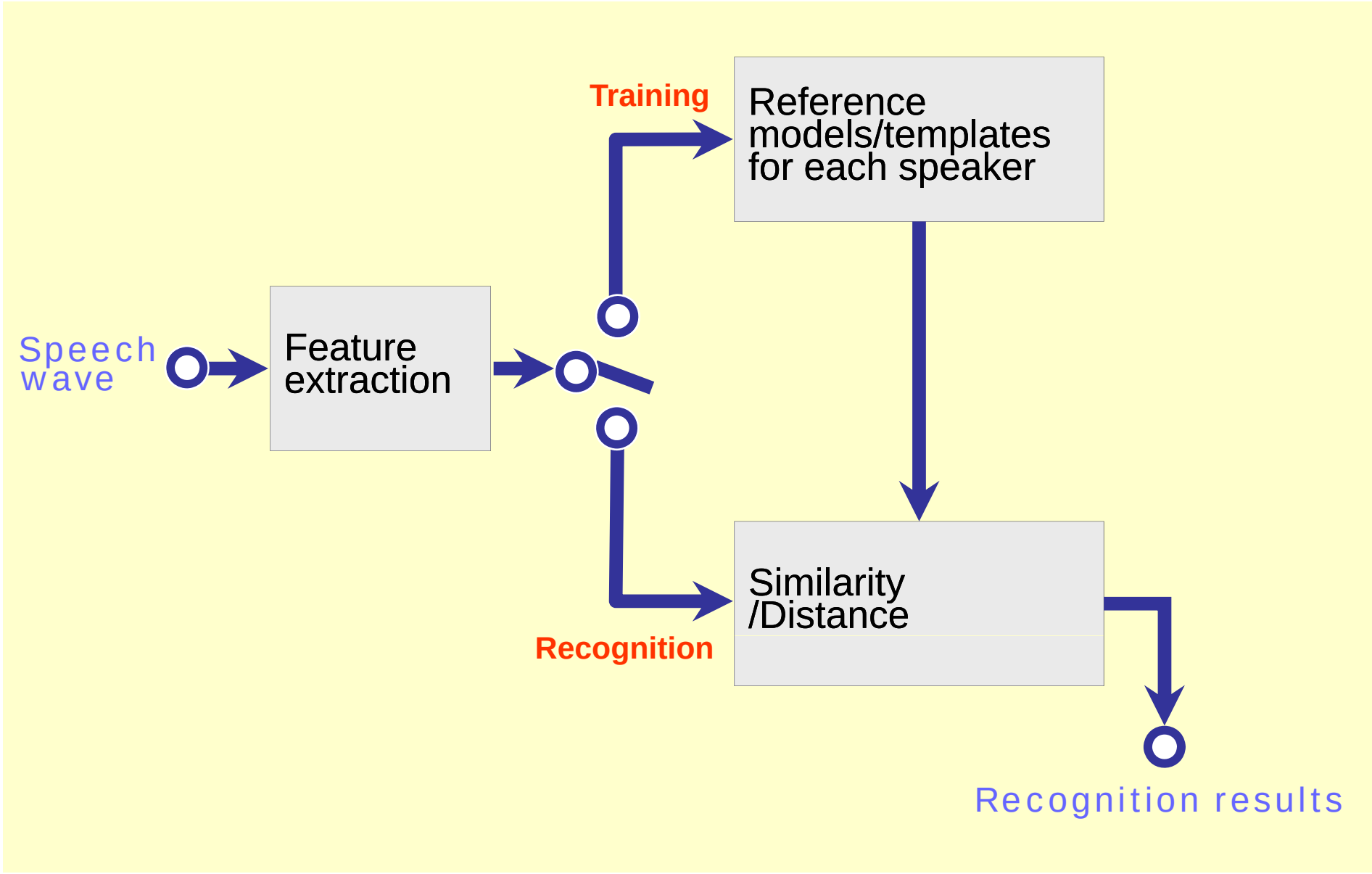
- Intersession variability (variability over time) of speech waves and spectra

⇒ Spectral/likelihood equalization (normalization)

Applications of speaker recognition technology

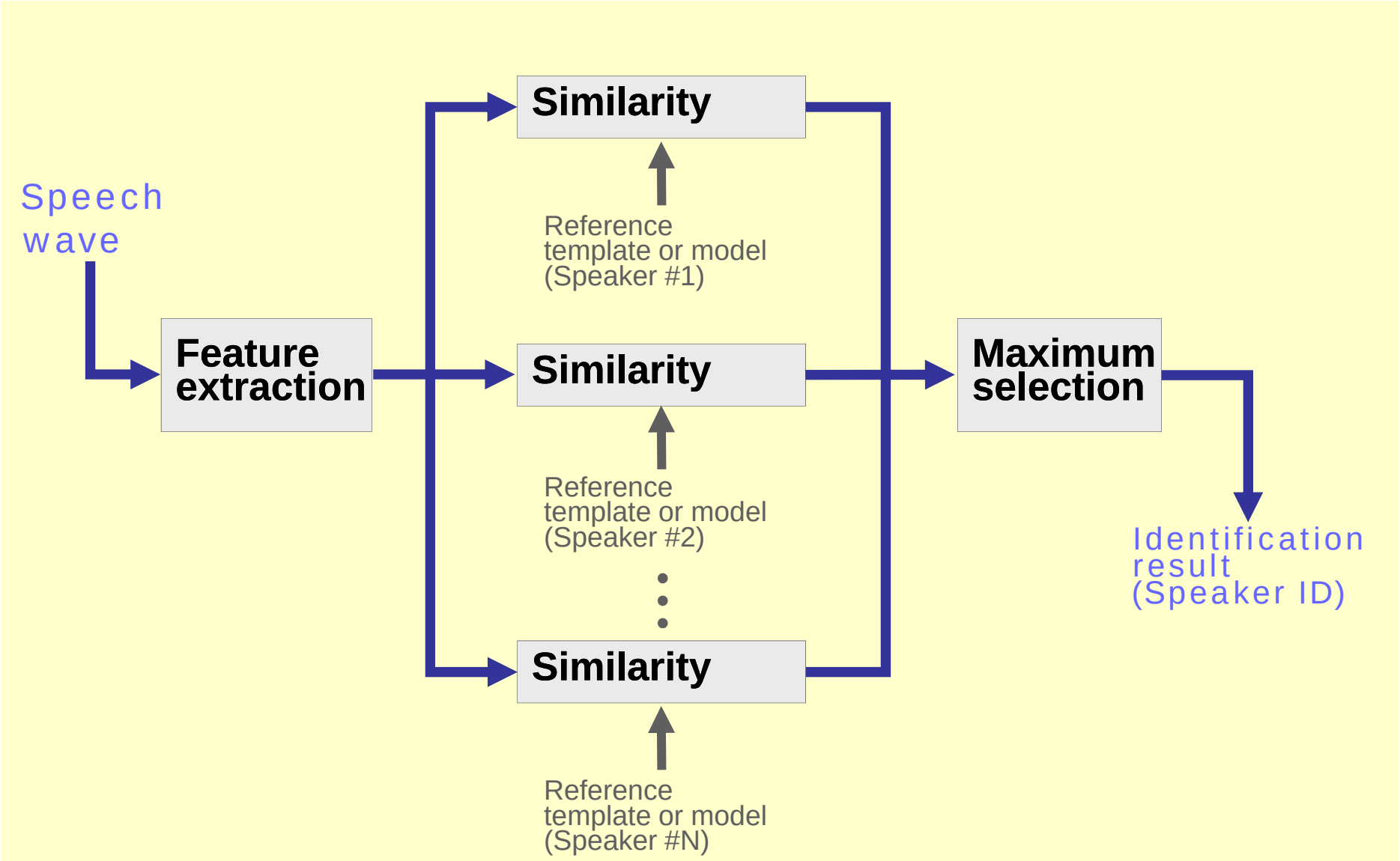
- **Access control**: For physical facilities, computer networks, websites and automated password reset services.
- **Transaction authentication**: For telephone banking and remote electronic and mobile purchases (e- and m-commerce).
- **Law enforcement**: Home-parole monitoring, prison call monitoring and corroborating aural/spectral inspections of voice samples for forensic analysis.
- **Speech data management**: Label incoming voice mail with speaker name for browsing and/or action. Annotate recorded meetings or video with speaker labels for quick indexing and filing.
- **Personalization**: Store and retrieve personal setting/preferences for multi-user site or device. Use speaker characteristics for directed advertisement or services.

Principal structure of speaker recognition systems



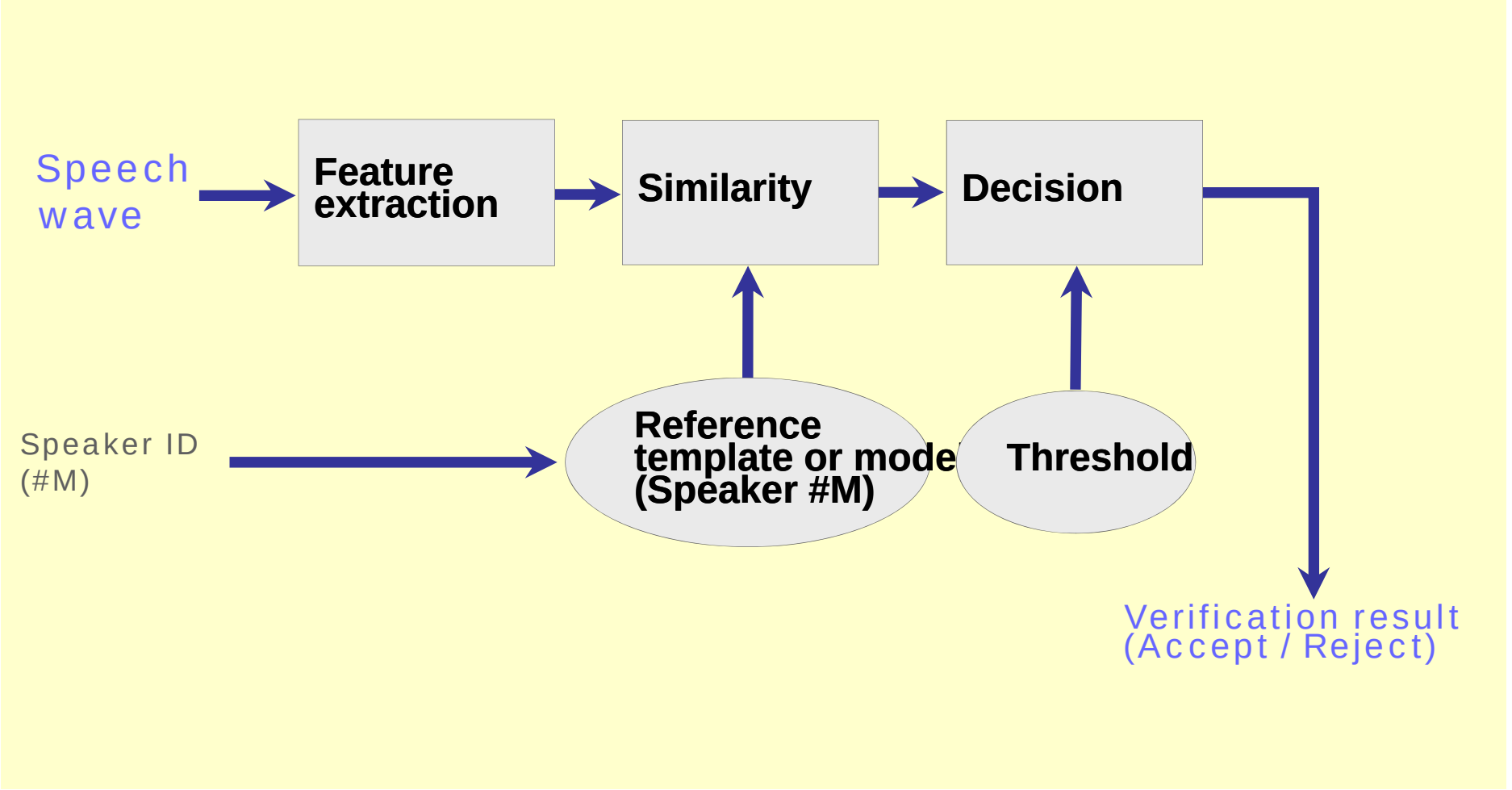
Basic structure of speaker recognition systems

(a) Speaker identification



Basic structure of speaker recognition systems

(b) Speaker verification



Past and future

- Speech recognition technology has made very significant progress in the past 50+ years with the help of computer technology.
- The majority of technological changes have been directed toward the purpose of increasing robustness of recognition.
- However, there still remain many unsolved problems.
- A much greater understanding of the human speech process is required before automatic speech recognition systems can approach human performance.
- Significant advances will come from extended knowledge processing in the framework of statistical pattern recognition.