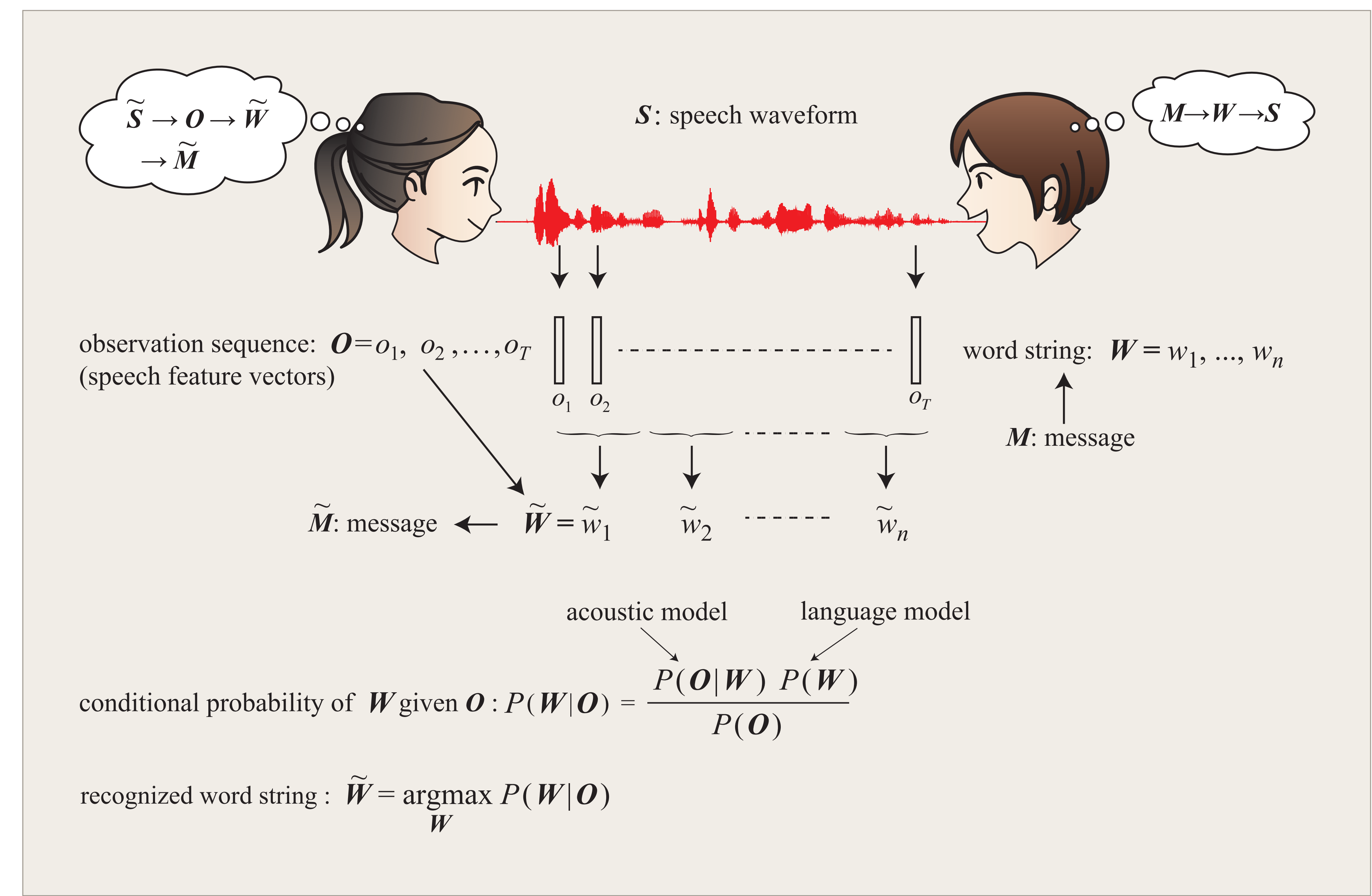


音声通信の統計的解釈

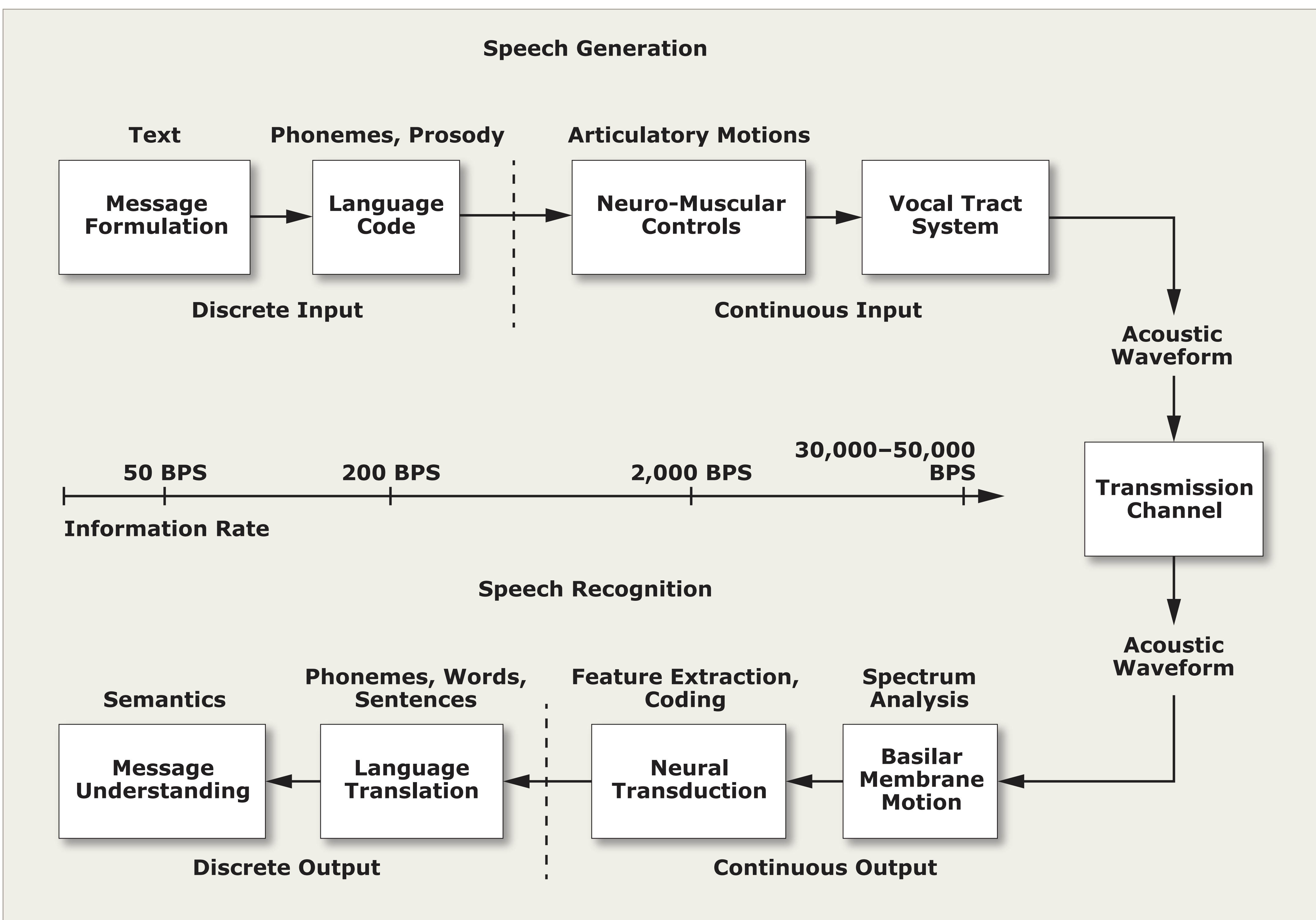
Statistical interpretation of voice communication



M : メッセージ (話者が伝えたい内容), W : 単語列 (発話), S : 音声波形, $\tilde{S}$: 聴き手が受け取った音声波形, O : 観測系列 (聴き手の内耳が生成した音声特徴時系列), $\tilde{W}$: 推定単語列, $\tilde{M}$: 推定メッセージ

M : message that the talker wants to transmit to the listener, W : word string formulated to express the message, S : speech waveform, $\tilde{S}$: speech waveform that the listener received, O : running spectrum provided by the inner ear of the listener, $\tilde{W}$: word string that the listener recognized, $\tilde{M}$: message that the listener recognized

音声認識の究極の目的は、音声波形 $\tilde{S}$ から話者のメッセージ M の推定値 $\tilde{M}$ を得ることであるともいえます。たとえば、部屋に人間とロボットが居るとして、人間が室温を下げて欲しいと思います。ロボットに「暑いですね。」と話したとします。その発話を聴いてロボットがエアコンの温度を調節して室温を下げてくれたならば、ロボットは人間のメッセージ (意図) を理解したといえることができます。ロボットにこのような対応をさせるためには、単に聴いた音声波形を「暑いですね。」という文に変換することができるだけではダメです。発話の字面の背後にある意図を理解することが必要です。そのような技術は音声理解 (speech understanding) と呼ばれています。現在、一般に音声認識と呼ばれている技術は音声理解までは行わず、観測系列 O から話者の発した単語列 W の推定値 $\tilde{W}$ を求めることを指しています。



Schematic representation of the human speech-communication process [Rabiner1981]

音声によるコミュニケーションの過程で生じる処理を追ってみましょう。話者は、まず、使う言語の意味構造にしたがってメッセージを組み立てます。次に、話者の話す言語に対応した言語符号 (文字列) に変換します。最後に、神経筋の活動が調整され、肺、声帯、声道の調音器官が適切な順番に動いて音声として聴こえる音波が生成されます。音声を聴く過程は末梢の調音器官に始まります。内耳の蝸牛では、非常に敏感で選択的なフィルタ群がスペクトル分析を行います。スペクトルと時間変化の情報は量子化され、符号化されて、脳の一次聴覚野に送られて、言語の記号として検知されます。脳の言語野では、言語の記号が解析され、メッセージの意味が理解されます。

Let us follow the processes which occur in the course of speech communication. A speaker wishing to convey a message must first formulate it in accordance with the semantic structure of the language. The conceptual representation must then be transformed into the corresponding linguistic encoding. Finally, neuromuscular activities must be coordinated so that the lungs, vocal cords, and articulatory organs of the vocal tract move in the proper sequence to produce the sound pressure wave which can be heard as speech. The hearing process begins in the peripheral auditory system, where spectral analysis is performed by the exquisitely sensitive and selective filter bank called the cochlea. Spectral and temporal information is quantized, encoded, and transmitted via the auditory nerve to the auditory cortex of the brain where symbol are detected, the linguistic code is deciphered and the semantic content extracted whereupon the message is understood. [Rabiner1981]

参考文献 References

[Rabiner1981] L.R. Rabiner and S.E. Levinson, "Isolated and Connected Word Recognition — Theory and Selected Applications," IEEE Transactions on Communications, COM-29 (5), pp.621–659, May 1981.

音声認識とは？
what is happening in
the speech recognition process?



ポスター PDF データ

音声通信の統計的解釈で定式化
formulated as statistical interpretation
of communication by speech