ELSEVIER

VOLUME TWELVE

ENCYCLOPEDIA OF LANGUAGE & LINGUISTICS

SECOND EDITION

EDITOR-IN-CHIEF
KEITH BROWN

CO-ORDINATING EDITORS
ANNE H. ANDERSON
LAURIE BAUER
MARGIE BERNS
GRAEME HIRST
JIM MILLER

ELL2

ENCYCLOPEDIA OF LANGUAGE & LINGUISTICS

SECOND EDITION

EDITOR-IN-CHIEF
KEITH BROWN

CO-ORDINATING EDITORS
ANNE H. ANDERSON
LAURIE BAUER
MARGIE BERNS
GRAEME HIRST
JIM MILLER

ELSEVIER

Amsterdam Boston Heidelberg London New York Oxford
Paris San Diego San Francisco Singapore Sydney Tokyo

Elsevier Ltd., The Boulevard, Langford Lane, Kidlington, Oxford, OX5 1GB, UK

First edition 2006

Library of Congress Control Number: 2005935119

A catalogue record for this book is available from the British Library

ISBN 0-08-044299-4 (set)

This book is printed on acid-free paper
Printed and bound in the United Kingdom

EXECUTIVE EDITORIAL BOARD/HONORARY EDITORIAL ADVISORY BOARD

HONORARY EDITORIAL ADVISORY BOARD MEMBERS

GUIDE TO USE OF THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of articles in alphabetical order. To help you realize the full potential of the material in the Encyclopedia we have provided several features to help you find the topic of your choice: an Alphabetical list of Articles, a Subject Classification, Cross-References and a Subject Index.

1. Alphabetical List of Articles

Your first point of reference will probably be the alphabetical list of articles. It provides a full alphabetical listing of all articles in the order they appear within the work. This list appears at the front of each volume, and will provide you with both the volume number and the page number of the article.

Alternatively, you may choose to browse through the work using the alphabetical order of the articles as your guide. To assist you in identifying your location within the Encyclopedia, a running head line indicates the current article.

You will also find 'dummy entries' for certain languages for which alternative language names exist within the alphabetical list of articles and body text.

For example, if you were attempting to locate material on the *Apalachee* language via the contents list, you would find the following:

Apalachee *See* Muskogean Languages.

The dummy entry directs you to the *Muskogean Languages* article.

If you were trying to locate the material by browsing through the text and you looked up *Apalachee*, you would find the following information provided in the dummy entry:

Apalachee *See:* Muskogean Languages.

2. Subject Classification

The subject classification is intended for use as a thematic guide to the contents of the Encyclopedia. It is divided by subject areas into 36 sections; most sections are further subdivided where appropriate. The sections and subdivisions appear alphabetically, as do the articles within each section. For quick reference, a list of the section headings and subheadings is provided at the start of the subject classification.

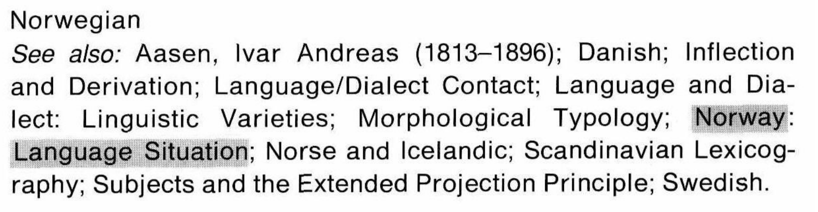
Every article in the encyclopedia is listed under at least one section, and a large number are also listed under one or more additional relevant sections. Biographical entries are an exception to this policy; they are listed only under biographies. Except for a very few cases, repeat entries have been avoided within sections, and a given article will appear only in the most appropriate subdivisions. Again, biographical entries are the main exception, with many linguists appearing in several subdivisions within biographies.

As explained in the introduction to the Encyclopedia, practical considerations necessitate that, of living linguists, only the older generation receive biographical entries. Those for members of the Encyclopedia's Honorary Editorial Advisory Board and Executive Editorial Board appear separately in Volume 1 and are not listed in the classified list of entries.

3. Cross-References

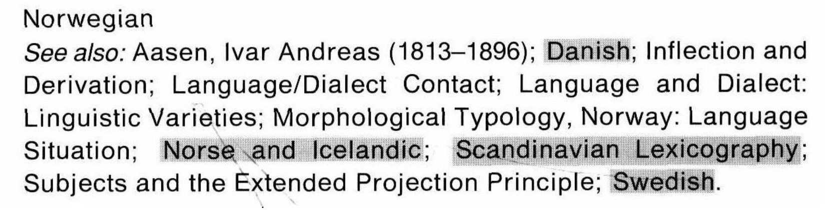
All of the articles in the Encyclopedia have been extensively cross-referenced. The cross-references, which appear at the end of each article, serve three different functions. For example, at the end of *Norwegian* article, cross-references are used:

1. to indicate if a topic is discussed in greater detail elsewhere



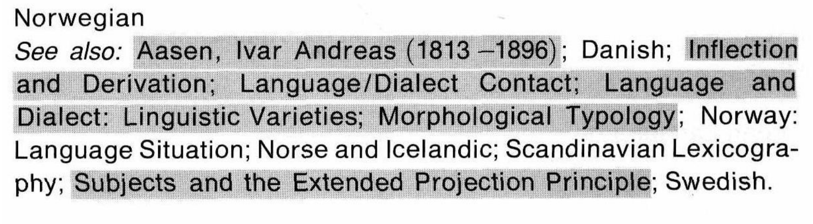
Norwegian
See also: Aasen, Ivar Andreas (1813–1896); Danish; Inflection and Derivation; Language/Dialect Contact; Language and Dialect: Linguistic Varieties; Morphological Typology; Norway: Language Situation; Norse and Icelandic; Scandinavian Lexicography; Subjects and the Extended Projection Principle; Swedish.

2. to draw the reader's attention to parallel discussions in other articles



Norwegian
See also: Aasen, Ivar Andreas (1813–1896); Danish; Inflection and Derivation; Language/Dialect Contact; Language and Dialect: Linguistic Varieties; Morphological Typology, Norway: Language Situation; Norse and Icelandic; Scandinavian Lexicography; Subjects and the Extended Projection Principle; Swedish.

3. to indicate material that broadens the discussion



Norwegian
See also: Aasen, Ivar Andreas (1813 –1896); Danish; Inflection and Derivation; Language/Dialect Contact; Language and Dialect: Linguistic Varieties; Morphological Typology; Norway: Language Situation; Norse and Icelandic; Scandinavian Lexicography; Subjects and the Extended Projection Principle; Swedish.

4. Subject Index

The index provides you with the page number where the material is located, and the index entries differentiate between material that is an entire article, part of an article, or data presented in a figure or table. Detailed notes are provided on the opening page of the index.

Other End Matter

In addition to the articles that form the main body of the Encyclopedia, there are 176 Ethnologue maps; a full list of contributors with contributor names, affiliations, and article titles; a List of Languages, and a Glossary. All of these appear in the last volume of the Encyclopedia.

ALPHABETICAL LIST OF ARTICLES
VOLUME 12

Note: Readers are urged to use the comprehensive name and subject indexes and the Classified List of Entries extensively, since the contents presented here represent only the broad framework of the *Encyclopedia.*

Speech Recognition: Statistical Methods
Speech Synthesis
Speech Synthesis: Perception and Comprehension of
Speech Synthesis: Prosody
Speech Technologies: Language Variation
Speech: Biological Basis
Spelling Reform
Spelling Reform Proposals: English
Spence, Thomas (1750–1814)
Spitzer, Leo (1887–1960)
Splinters
Spoken Discourse: Diachronic Aspects
Spoken Discourse: Morphophonological Aspects
Spoken Discourse: Types
Spoken Discourse: Word Order
Spoken Document Retrieval, Automatic
Spoken Language Production: Psycholinguistic Approach
Sports Broadcasting
Sri Lanka: Language Situation
St Christopher (Kitts) and Nevis: Language Situation
St Helena: Language Situation
St Lucia: Language Situation
St Pierre and Miquelon: Language Situation
St Vincent and the Grenadines: Language Situation
Standard and Dialect Vocabulary
Standardization
Stang, Christian S. (b. 1900)
States of the Glottis
Statistical Testing of Variation
Statistics
Steele, Joshua (1700–1791)
Steinthal, Heymann (1823–1899)
Stemming
Stephanus, Henricus (1531–1598)
Stephanus, Robertus (ca. 1503–1559)
Stereotype Semantics
Stokes, Whitley (1830–1909)
Stokoe,William (1919–2000)
Storm, Johan (1836–1920)
Strachan, John (1862–1907)
Strawson, Peter Frederick (b. 1919)
Stroop Effect in Language
Structuralism
Structuralism in Anthropology
Structuralist Phonology: Prague School
Structure and Structuralism: Semiotic Approaches
Structures and Processes
Sturge-Weber-Dimitri Syndrome and Language
Sturtevant, Edgar Howard (1875–1952)
Style
Style and Style Shifting
Stylistics
Stylistics, Cognitive
Stylistics, Feminist
Stylistics: Corpus Approaches
Stylistics: Pedagogical Applications
Stylistics: Pragmatic Approaches
Subcultures and Countercultures
Subgrouping Methodology
Subjective and Objective Measures of Variation
Subjects and the Extended Projection Principle
Subordination
Subordination in Spoken Discourse
Subtitling
Subtraction

Sudan: Language Situation
Sumerian
Summarization of Text, Automatic
Summer Institute of Linguistics
Sumu *See* Misumalpan
Suppletion
Suriname: Language Situation
Suzhou Wu *See* Chinese
Suzuki Akira (1764–1837)
Swadesh, Morris (1909–1967)
Swahili
Swales, John (b. 1938)
Swaziland: Language Situation
Sweden: Language Situation
Swedish
Sweet, Henry (1845–1912)
Switch-Reference
Switzerland: Language Situation
Syllabic Constituents
Syllable: Phonology
Syllable: Typology
Symbolic Computational Linguistics: Overview
Syncategoremata
Synchronic and Diachronic Variation
Syncretism
Synesthesia
Synesthesia and Language
Synonymy
Syntactic Constructions
Syntactic Development
Syntactic Features and Feature Structures
Syntactic Variables and Variable-free Syntax
Syntactic Variation
Syntax of Words
Syntax: Optimality Theory
Syntax-Phonology Interface
Syntax-Pragmatics Interface: Overview
Syntax-Semantics Interface
Syria: Language Situation
Syriac
Systematicity
Systemic Theory
Szemerényi, Oswald J. L. (b. 1913–1996)
Taboo
Taboo Words
Taboo, Euphemism, and Political Correctness
Taboo: Verbal Practices
Tacit Knowledge
Tadadjeu, Maurice (b. 1950)
Tadzhik *See* Tajik Persian
Tagalog
Tagliavini, Carlo (1903–1982)
Tagmemics
Tahitian
Tai Languages
Taiwan: Language Situation
Tajik Persian
Tajikistan: Language Situation
Talk Shows
Tamambo
Tamil
Tamil Linguistic Tradition
Tannen, Deborah (b. 1945)
Tanzania: Language Situation
Tariana
Tarski, Alfred (1902–1983)
Tatar
Teacher Preparation: Second Language
Teaching of Minority Languages
Teaching Technologies: Second Language
Technical Translation
Teenagers, Variation, and Young People's Culture
Telephone Talk
Television: Language
Telugu
Template Morphology
Temporal Logic
Tense
Tense and Time: Philosophical Aspects
Tense, Mood, Aspect: Overview
Terminology and Terminological Databases
Terminology, Term Banks and Termbases for Translation
Terracini, Aron Benvenuto (1886–1968)
Tesnière, Lucien Valerius (1893–1954)
Testimony
Text and Text Analysis
Text Formatting
Text Mining
Text Retrieval Conference and Message Understanding Conference
Text World Theory
Texts: Semiotic Theory
Thailand: Language Situation
Thelwall, John (1764–1834)
Thematic Structure
Thematics
Theme in Text
Thesauruses
Thetic-Categorical Distinction
Think-Aloud Protocols
Third Language Acquisition
Thomas of Erfurt (fl. 1300)
Thomas, John Jacob (ca. 1840–1889)
Thompson, Laurence Cassius (b. 1926)
Thomsen, Vilhelm Ludvig Peter (1824–1927)
Thought and Language: Philosophical Aspects
Thought Experiments
Thumb, Albert (1865–1915)

Thurneysen, Rudolf (1857–1940)
Tibet: Scripts
Tibetan
Tigrinya
Tillamook *See* Salishan Languages
Tilly, William (1860–1935)
Timor Lorosae: Language Situation
Tinbergen, Niko (1917–1988)
Tiwi
Tobler, Adolf (1835–1910)
Tocharian
Toda
Todo, Akiyaso (b. 1915)
Tofan *See* Turkic Languages
Togo: Language Situation
Tohono O'odham
Tok Pisin
Tokelau: Language Situation
Tokieda, Motoki (1900–1967)
Toma, Peter (b. 1924)
Tone in Connected Discourse
Tone: Neurophonetics
Tone: Phonology
Tonga: Language Situation
Tooke, John Horne (1736–1812)
Topic and Comment
Topic Chaining Constructions
Toporov, Vladimir Nikolaevich (b. 1928)

Speech Recognition: Statistical Methods

L R Rabiner, Rutgers University, New Brunswick, NJ, USA and University of California, Santa Barbara, CA, USA
B-H Juang, Georgia Institute of Technology, Atlanta, GA, USA

Introduction

The goal of getting a machine to understand fluently spoken speech and respond in a natural voice has been driving speech research for more than 50 years. Although the personification of an intelligent machine such as HAL in the movie *2001, A Space Odyssey*, or R2D2 in the *Star Wars* series, has been around for more than 35 years, we are still not yet at the point where machines reliably understand fluent speech, spoken by anyone, and in any acoustic environment. In spite of the remaining technical problems that need to be solved, the fields of automatic speech recognition and understanding have made tremendous advances and the technology is now readily available and used on a day-to-day basis in a number of applications and services – especially those conducted over the public-switched telephone network (PSTN) (Cox *et al.*, 2000). This article aims at reviewing the technology that has made these applications possible.

Speech recognition and language understanding are two major research thrusts that have traditionally been approached as problems in linguistics and acoustic phonetics, where a range of acoustic phonetic knowledge has been brought to bear on the problem with remarkably little success. In this article, however, we focus on statistical methods for speech and language processing, where the knowledge about a speech signal and the language that it expresses, together with practical uses of the knowledge, is developed from actual realizations of speech data through a well-defined mathematical and statistical formalism. We review how the statistical methods are used for speech recognition and language understanding, show current performance on a number of task-specific applications and services, and discuss the challenges that remain to be solved before the technology becomes ubiquitous.

The Speech Advantage

There are fundamentally three major reasons why so much research and effort has gone into the problem of trying to teach machines to recognize and understand fluent speech, and these are the following:

- Cost reduction. Among the earliest goals for speech recognition systems was to replace humans performing certain simple tasks with automated machines, thereby reducing labor expenses while still providing customers with a natural and convenient way to access information and services. One simple example of a cost reduction system was the Voice Recognition Call Processing (VRCP) system introduced by AT&T in 1992 (Roe *et al.*, 1996), which essentially automated so-called operator-assisted calls, such as person-to-person calls, reverse-billing calls, third-party billing calls, collect calls (by far the most common class of such calls), and operator-assisted calls. The resulting automation eliminated about 6600 jobs, while providing a quality of service that matched or exceeded that provided by the live attendants, saving AT&T on the order of $300 million per year.
- New revenue opportunities. Speech recognition and understanding systems enabled service providers to have a 24/7 high-quality customer care automation capability, without the need for access to information by keyboard or touch-tone button pushes. An example of such a service was the How May I Help You (HMIHY)© service introduced by AT&T late in 2000 (Gorin *et al.*, 1996), which automated the customer care for AT&T Consumer Services. This system will be discussed further in the section on speech understanding. A second example of such a service was the NTT ANSER service for voice banking in Japan [Sugamura *et al.*, 1994], which enabled Japanese banking customers to access bank account records from an ordinary telephone without having to go to the bank. (Of course, today we utilize the Internet for such information, but in 1981, when this system was introduced, the only way to access such records was a physical trip to the bank and a wait in lines to speak to a banking clerk.)
- Customer retention. Speech recognition provides the potential for personalized services based on customer preferences, and thereby the potential to improve the customer experience. A trivial example of such a service is the voice-controlled automotive environment that recognizes the identity of the driver from voice commands and adjusts the automobile's features (seat position, radio station, mirror positions, etc.) to suit the customer's preference (which is established in an enrollment session).

The Speech Dialog Circle

When we consider the problem of communicating with a machine, we must consider the cycle of events that occurs between a spoken utterance (as part of a dialog between a person and a machine) and the response to that utterance from the machine. **Figure 1** shows such a sequence of events, which is often referred to as the speech dialog circle, using an example in the telecommunications context.

The customer initially makes a request by speaking an utterance that is sent to a machine, which attempts to recognize, on a word-by-word basis, the spoken speech. The process of recognizing the words in the speech is called automatic speech recognition (ASR) and its output is an orthographic representation of the recognized spoken input. The ASR process will be discussed in the next section. Next the spoken words are analyzed by a spoken language understanding (SLU) module, which attempts to attribute meaning to the spoken words. The meaning that is attributed is in the context of the task being handled by the speech dialog system. (What is described here is traditionally referred to as a limited domain understanding system or application.) Once meaning has been determined, the dialog management (DM) module examines the state of the dialog according to a prescribed operational workflow and determines the course of action that would be most appropriate to take. The action may be as simple as a request for further information or confirmation of an action that is taken. Thus if there were confusion as to how best to proceed, a text query would be generated by the spoken language generation module to hopefully clarify the meaning and help determine what to do next. The query text is then sent to the final module, the text-to-speech synthesis (TTS) module, and then converted into intelligible and highly natural speech, which is sent to the customer who decides what to say next based on what action was taken, or based on previous dialogs with the machine. All of the modules in the speech dialog circle can be 'data-driven' in both the learning and active use phases, as indicated by the central Data block in **Figure 1**.

A typical task scenario, e.g., booking an airline reservation, requires navigating the speech dialog circle many times – each time being referred to as one 'turn' – to complete a transaction. (The average number of turns a machine takes to complete a prescribed task is a measure of the effectiveness of the machine in many applications.) Hopefully, each time through the dialog circle enables the customer to get closer to the desired action either via proper understanding of the spoken request or via a series of clarification steps. The speech dialog circle is a powerful concept in modern speech recognition and understanding systems, and is at the heart of most speech understanding systems that are in use today.

Basic ASR Formulation

The goal of an ASR system is to accurately and efficiently convert a speech signal into a text message transcription of the spoken words, independent of the device used to record the speech (i.e., the

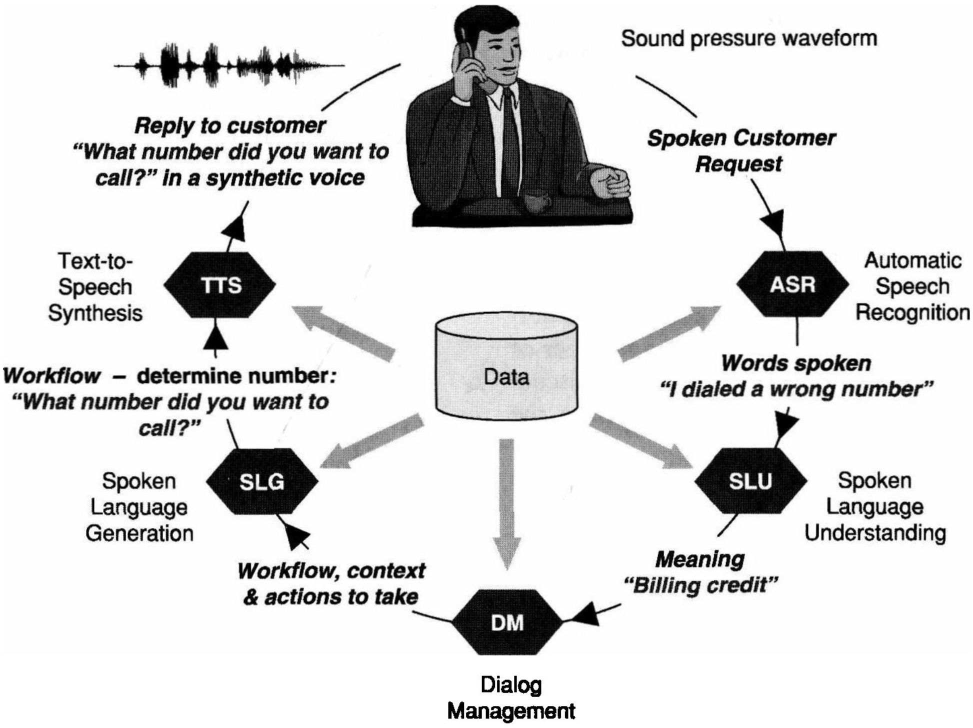

Sound pressure waveform
Spoken Customer Request
ASR
Automatic Speech Recognition
Words spoken
"I dialed a wrong number"
SLU
Spoken Language Understanding
Meaning
"Billing credit"
DM
Dialog Management
Workflow, context & actions to take
SLG
Spoken Language Generation
Workflow – determine number: "What number did you want to call?"
TTS
Text-to-Speech Synthesis
Reply to customer
"What number did you want to call?" in a synthetic voice
Data

Figure 1 The conventional speech dialog circle.

transducer or microphone), the speaker, or the environment. A simple model of the speech generation process, as used to convey a speaker's intention is shown in **Figure 2**.

It is assumed that the speaker decides what to say and then embeds the concept in a sentence, W, which is a sequence of words (possibly with pauses and other acoustic events such as uh's, um's, er's, etc.). The speech production mechanisms then produce a speech waveform, $s(n)$, which embodies the words of W as well as the extraneous sounds and pauses in the spoken input. A conventional automatic speech recognizer attempts to decode the speech, $s(n)$, into the best estimate of the sentence, $\hat{W}$, using a two-step process, as shown in **Figure 3**.

The first step in the process is to convert the speech signal, $s(n)$, into a sequence of spectral feature vectors, X, where the feature vectors are measured every 10 ms (or so) throughout the duration of the speech signal. The second step in the process is to use a syntactic decoder to generate every possible valid sentence (as a sequence of orthographic representations) in the task language, and to evaluate the score (i.e., the *a posteriori* probability of the word string given the realized acoustic signal as measured by the feature vector) for each such string, choosing as the recognized string, $\hat{W}$, the one with the highest score. This is the so-called maximum *a posteriori* probability (MAP) decision principle, originally suggested by Bayes. Additional linguistic processing can be done to try to determine side information about the speaker, such as the speaker's intention, as indicated in **Figure 3**.

Mathematically, we seek to find the string $\hat{W}$ that maximizes the *a posteriori* probability of that string, when given the measured feature vector X, i.e.,

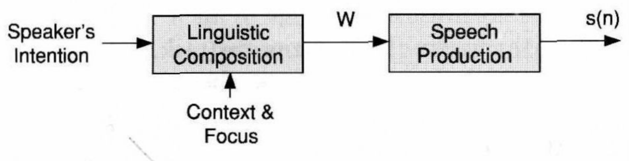


Figure 2 Model of spoken speech.

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

Using Bayes Law, we can rewrite this expression as:

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

Thus, calculation of the *a posteriori* probability is decomposed into two main components, one that defines the *a priori* probability of a word sequence W, $P(W)$, and the other the likelihood of the word string W in producing the measured feature vector, $P(X|W)$. (We disregard the denominator term, $P(X)$, since it is independent of the unknown W). The latter is referred to as the acoustic model, $P_A(X|W)$, and the former the language model, $P_L(W)$ (Rabiner *et al.*, 1996; Gauvain and Lamel, 2003). We note that these quantities are not given directly, but instead are usually estimated or inferred from a set of training data that have been labeled by a knowledge source, i.e., a human expert. The decoding equation is then rewritten as:

$$\hat{W} = \arg\max_{W} P_A(X|W)P_L(W)$$

We explicitly write the sequence of feature vectors (the acoustic observations) as:

$$X = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$$

where the speech signal duration is N frames (or N times 10 ms when the frame shift is 10 ms). Similarly, we explicitly write the optimally decoded word sequence as:

$$\hat{W} = w_1 w_2 \ldots w_M$$

where there are M words in the decoded string. The above decoding equation defines the fundamental statistical approach to the problem of automatic speech recognition.

It can be seen that there are three steps to the basic ASR formulation, namely:

- Step 1: acoustic modeling for assigning probabilities to acoustic (spectral) realizations of a sequence

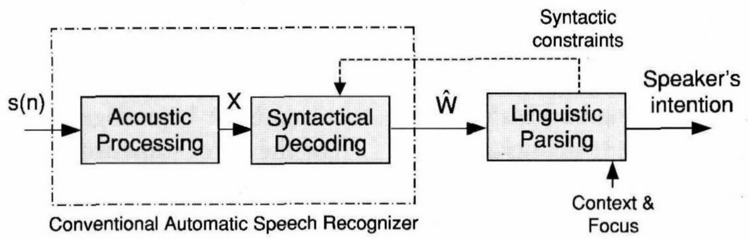


Figure 3 ASR decoder from speech to sentence.

of words. For this step we use a statistical model (called the hidden Markov model or HMM) of the acoustic signals of either individual words or subword units (e.g., phonemes) to compute the quantity $P_A(X|W)$. We train the acoustic models from a training set of speech utterances, which have been appropriately labeled to establish the statistical relationship between X and W.

- Step 2: language modeling for assigning probabilities, $P_L(W)$, to sequences of words that form valid sentences in the language and are consistent with the recognition task being performed. We train such language models from generic text sequences, or from transcriptions of task-specific dialogues. (Note that a deterministic grammar, as is used in many simple tasks, can be considered a degenerate form of a statistical language model. The 'coverage' of a deterministic grammar is the set of permissible word sequences, i.e., expressions that are deemed legitimate.)
- Step 3: hypothesis search whereby we find the word sequence with the maximum *a posteriori* probability by searching through all possible word sequences in the language.

In step 1, acoustic modeling (Young, 1996; Rabiner *et al.*, 1986), we train a set of acoustic models for the words or sounds of the language by learning the statistics of the acoustic features, X, for each word or sound, from a speech training set, where we compute the variability of the acoustic features during the production of the words or sounds, as represented by the models. For large vocabulary tasks, it is impractical to create a separate acoustic model for every possible word in the language since it requires far too much training data to measure the variability in every possible context. Instead, we train a set of about 50 acoustic-phonetic subword models for the approximately 50 phonemes in the English language, and construct a model for a word by concatenating (stringing together sequentially) the models for the constituent subword sounds in the word, as defined in a word lexicon or dictionary, where multiple pronunciations are allowed). Similarly, we build sentences (sequences of words) by concatenating word models. Since the actual pronunciation of a phoneme may be influenced by neighboring phonemes (those occurring before and after the phoneme), the set of so-called context-dependent phoneme models are often used as the speech models, as long as sufficient data are collected for proper training of these models.

In step 2, the language model (Jelinek, 1997; Rosenfeld, 2000) describes the probability of a sequence of words that form a valid sentence in the task language. A simple statistical method works well, based on a Markovian assumption, namely that the probability of a word in a sentence is conditioned on only the previous N–1 words, namely an N-gram language model, of the form:

$$\begin{aligned} P_L(W) &= P_L(w_1, w_2, \ldots, w_M) \\ &= \prod_{m=1}^{M} P_L(w_m|w_{m-1}, w_{m-2}, \ldots, w_{m-N+1}) \end{aligned}$$

where $P_L(w_m|w_{m-1}, w_{m-2}, \ldots, w_{m-N+1})$ is estimated by simply counting up the relative frequencies of N-tuples in a large corpus of text.

In step 3, the search problem (Ney, 1984; Paul, 2001) is one of searching the space of all valid sound sequences, conditioned on the word grammar, the language syntax, and the task constraints, to find the word sequence with the maximum likelihood. The size of the search space can be astronomically large and take inordinate amounts of computing power to solve by heuristic methods. The use of methods from the field of Finite State Automata Theory provide finite state networks (FSNs) (Mohri, 1997), along with the associated search policy based on dynamic programming, that reduce the computational burden by orders of magnitude, thereby enabling exact solutions in computationally feasible times, for large speech recognition problems.

Development of a Speech Recognition System for a Task or an Application

Before going into more detail on the various aspects of the process of automatic speech recognition by machine, we review the three steps that must occur in order to define, train, and build an ASR system (Juang *et al.*, 1995; Kam and Helander, 1997). These steps are the following:

- Step 1: choose the recognition task. Specify the word vocabulary for the task, the set of units that will be modeled by the acoustic models (e.g., whole words, phonemes, etc.), the word pronunciation lexicon (or dictionary) that describes the variations in word pronunciation, the task syntax (grammar), and the task semantics. By way of example, for a simple speech recognition system capable of recognizing a spoken credit card number using isolated digits (i.e., single digits spoken one at a time), the sounds to be recognized are either whole words or the set of subword units that appear in the digits /zero/ to /nine/ plus the word /oh/. The word vocabulary is the set of 11 digits. The task syntax allows any single digit to be spoken, and the task

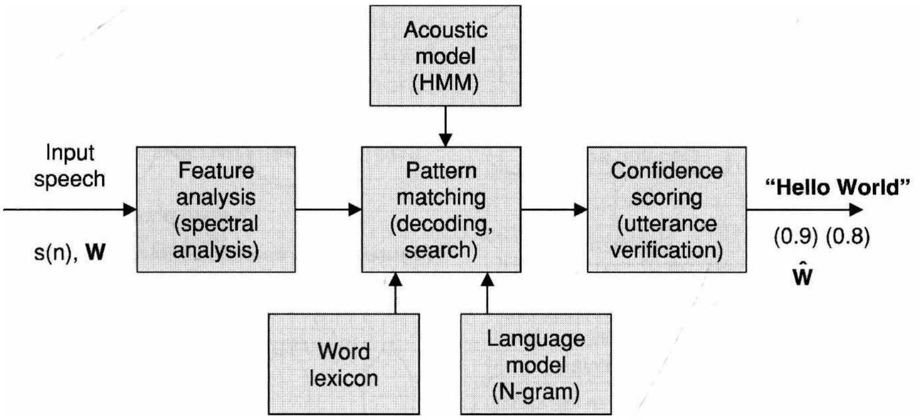

Acoustic model (HMM)
Input speech
s(n), W
Feature analysis (spectral analysis)
Pattern matching (decoding, search)
Confidence scoring (utterance verification)
"Hello World"
(0.9) (0.8)
$\hat{W}$
Word lexicon
Language model (N-gram)

Figure 4 Framework of ASR system.

semantics specify that a sequence of isolated digits must form a valid credit card code for identifying the user.

- Step 2: train the models. Create a method for building acoustic word models (or subword models) from a labeled speech training data set of multiple occurrences of each of the vocabulary words by one or more speakers. We also must use a text training data set to create a word lexicon (dictionary) describing the ways that each word can be pronounced (assuming we are using subword units to characterize individual words), a word grammar (or language model) that describes how words are concatenated to form valid sentences (i.e., credit card numbers), and finally a task grammar that describes which valid word strings are meaningful in the task application (e.g., valid credit card numbers).
- Step 3: evaluate recognizer performance. We need to determine the word error rate and the task error rate for the recognizer on the desired task. For an isolated digit recognition task, the word error rate is just the isolated digit error rate, whereas the task error rate would be the number of credit card errors that lead to misidentification of the user. Evaluation of the recognizer performance often includes an analysis of the types of recognition errors made by the system. This analysis can lead to revision of the task in a number of ways, ranging from changing the vocabulary words or the grammar (i.e., to eliminate highly confusable words) to the use of word spotting, as opposed to word transcription. As an example, in limited vocabulary applications, if the recognizer encounters frequent confusions between words like 'freight' and 'flight,' it may be advisable to change 'freight' to 'cargo' to maximize its distinction from 'flight.' Revision of the task grammar often becomes necessary if the recognizer experiences substantial amounts of what is called 'out of grammar' (OOG) utterances, namely the use of words and phrases that are not directly included in the task vocabulary (ISCA, 2001).

The Speech Recognition Process

In this section, we provide some technical aspects of a typical speech recognition system. **Figure 4** shows a block diagram of a speech recognizer that follows the Bayesian framework discussed above.

The recognizer consists of three processing steps, namely feature analysis, pattern matching, and confidence scoring, along with three trained databases, the set of acoustic models, the word lexicon, and the language model. In this section, we briefly describe each of the processing steps and each of the trained model databases.

Feature Analysis

The goal of feature analysis is to extract a set of salient features that characterize the spectral properties of the various speech sounds (the subword units) and that can be efficiently measured. The 'standard' feature set for speech recognition is a set of mel-frequency cepstral coefficients (MFCCs) (which perceptually match some of the characteristics of the spectral analysis done in the human auditory system) (Davis and Mermelstein, 1980), along with the first- and second-order derivatives of these features. Typically about 13 MFCCs and their first and second derivatives (Furai, 1981) are calculated every 10 ms, leading to a spectral vector with 39 coefficients every 10 ms. A block diagram of a typical feature analysis process is shown in **Figure 5**.

The speech signal is sampled and quantized, pre-emphasized by a first-order (highpass) digital filter with pre-emphasis factor α (to reduce the influence of glottal coupling and lip radiation on the estimated

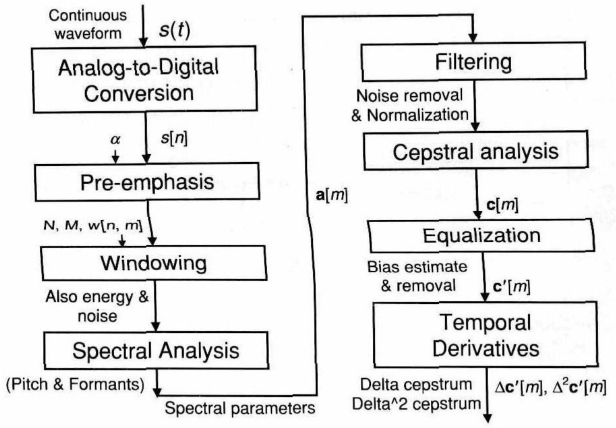


Figure 5 Block diagram of feature analysis computation.

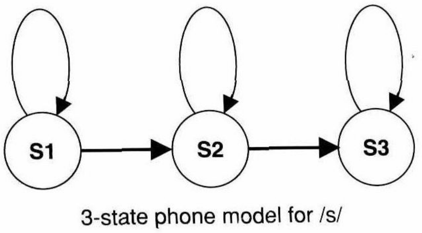


Figure 6 Three-state HMM for the sound /s/.

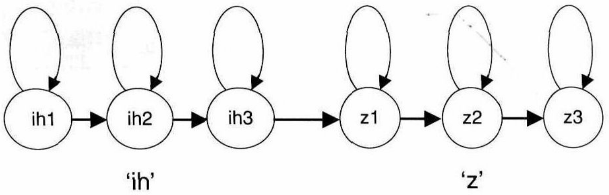


Figure 7 Concatenated model for the word 'is.'

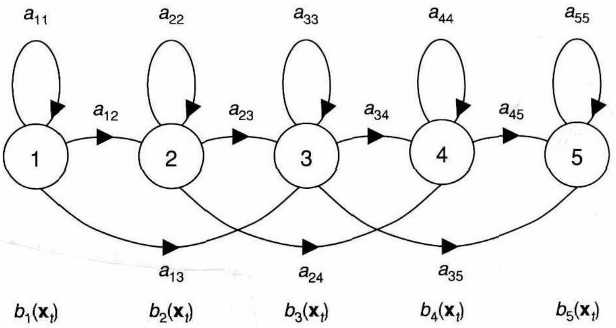


Figure 8 HMM for whole word model with five states.

vocal tract characteristics), segmented into frames, windowed, and then a spectral analysis is performed using a fast Fourier transform (FFT) (Rabiner and Gold, 1975) or linear predictive coding (LPC) method (Atal and Hanauer, 1971; Markel and Gray, 1976). The frequency conversion from a linear frequency scale to a mel frequency scale is performed in the filtering block, followed by cepstral analysis yielding the MFCCs (Davis and Mermelstein, 1980), equalization to remove any bias and to normalize the cepstral coefficients (Rahim and Juang, 1996), and finally the computation of first- and second-order (via temporal derivative) MFCCs is made, completing the feature extraction process.

Acoustic Models

The goal of acoustic modeling is to characterize the statistical variability of the feature set determined above for each of the basic sounds (or words) of the language. Acoustic modeling uses probability measures to characterize sound realization using statistical models. A statistical method, known as the hidden Markov model (HMM) (Levinson *et al.*, 1983; Ferguson, 1980; Rabiner, 1989; Rabiner and Juang, 1985), is used to model the spectral variability of each of the basic sounds of the language using a mixture density Gaussian distribution (Juang *et al.*, 1986; Juang, 1985), which is optimally aligned with a speech training set and iteratively updated and improved (the means, variances, and mixture gains are iteratively updated) until an optimal alignment and match is achieved.

Figure 6 shows a simple three-state HMM for modeling the subword unit /s/ as spoken at the beginning of the word /six/. Each HMM state is characterized by a probability density function (usually a mixture Gaussian density) that characterizes the statistical behavior of the feature vectors at the beginning (state s1), middle (state s2), and end (state s3) of the sound /s/. In order to train the HMM for each subword unit, we use a labeled training set of words and sentences and utilize an efficient training procedure known as the Baum-Welch algorithm (Rabiner, 1989; Baum, 1972; Baum *et al.*, 1970) to align each of the various subword units with the spoken inputs, and then estimate the appropriate means, covariances, and mixture gains for the distributions in each subword unit state. The algorithm is a hill-climbing algorithm and is iterated until a stable alignment of subword unit models and speech is obtained, enabling the creation of stable models for each subword unit.

Figure 7 shows how a simple two-sound word, 'is,' which consists of the sounds /ih/ and /z/, is created by concatenating the models (Lee, 1989) for the /ih/ sound with the model for the /z/ sound, thereby creating a six-state model for the word 'is.'

Figure 8 shows how an HMM can be used to characterize a whole-word model (Lee *et al.*, 1989). In this