

Speech Technology: Past, Present And Future

TORBJØRN SVENDSEN



Torbjørn Svendsen (48) received his *Siv. Ing. (MSc)* and *Dr. Ing. (PhD)* degrees in Electrical Engineering from the Norwegian University of Science and Technology (NTNU) in 1979 and 1985, respectively. The topic of the theses was speech coding. 1981–1988 he was a scientist at SINTEF DELAB working primarily on speech coding and recognition. In 1988 he joined the Dept. of Telecommunications at NTNU, where he currently is professor. He has been visiting professor at Griffith University and Queensland University of Technology, and has had several research visits to AT&T Bell Laboratories and AT&T Labs. Dr. Svendsen's main field of research is speech processing, encompassing speech recognition, speech coding and speech synthesis. He is a member of IEEE, the European Speech Communication Association (ESCA) and NORSIG.

torbjorn@tele.ntnu.no

Speech technology is intended to enable human-machine communication by voice. This includes making computers able to understand human speech, *speech recognition*, to produce intelligible and natural sounding speech, *speech synthesis*, and to determine who is speaking, *speaker recognition*. Unlike many other technology areas, speech technology is closely linked to an understanding of basic human properties. Without knowledge of the fundamentals of human speech production and perception, progress in speech technology will be difficult to achieve. In this paper we start out with an overview of human speech communication before describing the main trends in the development of speech technology, from the early efforts of mimicking human speech production up till today's, mainly data-driven statistical approaches. Although the progress in speech technology performance has been great over the last 25 years, the technology is far from perfect. Many useful products and services employing speech technology exist today, but significant research issues need to be resolved before computers can approach the performance of human listeners and talkers.

Introduction

The capacity for speech communication is a trait that distinguishes man from other species. Consequently, speech has also fascinated humans for centuries. In ancient times, objects that seemingly could produce and understand speech were attributed with having supernatural powers. One such phenomenon was the Greek oracles, which would give cryptic answers and predictions to the questions presented. Today, we know that the voice of the oracle emanated from a person hidden inside the oracle, speaking into a tube that led into the oracle's head.

In more recent times, speech enabled objects are still desired, although the supernatural notions are long abandoned. Speech is a simple means of communication for humans, and has the advantage that communication can be undertaken while your hands and eyes are busy with other tasks. Speaking and listening does not require that you learn motor skills like using a keyboard, and does not even require that you are able to read and write.

Speech technology can enable us to communicate with machinery by voice. This includes the technology that can enable machines to produce intelligible and natural sounding speech: *speech synthesis*, and the technology for recognizing or understanding what we speak: *speech recognition*. Combining these technologies is necessary in order to produce *spoken dialogue systems*, systems where the interaction between user and machine is based on speech.

In spite of intensive research efforts over nearly half a century, making truly speech enabled machines remains an elusive goal. Although the technology has come a long way since the first efforts at producing electronic machinery for this task, the current state-of-the-art systems can be

described as being at an early adolescent evolutionary stage. The simple tasks are well mastered. More complex tasks are also mastered when the circumstances are favorable and the environment well controlled, but the performance tends to fail when this is not the case. The technology that will enable spoken human-machine communication to be as effortless and reliable as spoken communication between humans, for all acoustic environments, topics of conversation and speaker accents is still far from reality. Indeed, until we have a deeper and more complete understanding of the human speech communication process, and of how technology can be applied to speech communication, speech technology will only be successful in restricted application domains.

What is it that makes speech communication such a difficult task for machines? Most humans master the process of speech communication without much drama, although learning to speak and to understand speech can take many years. Once learned, the human capacity for e.g. recognizing speech in noisy environments is quite amazing. Near perfect speech recognition in environments where the noise level exceeds the speech level is something most people do without too much effort. An automatic speech recognizer on the other hand will be rendered nearly useless if the SNR drops below 10–15 dB. Changes in speech patterns, topic shifts etc. are also handled without problems by humans while the machines have huge difficulties with situations that deviate from a well-defined and well-known setting.

In this paper, we will take a look at some of the main technology development trends in different speech technology areas. We will begin with a short description of the basics of speech communications, speech production and perception

before moving to the technologies that aim to enable machines to speak, and to recognize (and understand) speech.

Speech Communication

Let us take a look at the steps involved in speech communication. In Figure 1, a communication-theory inspired interpretation of the steps involved in the speech communication chain is depicted [16]. The initial step is generation in our mind of a notion or an idea that we wish to convey. This is represented as an output, M , from a message source, with an associated probability, $P(M)$. The message will need to be formulated in words, a transformation that is dependent on the semantics, syntax and grammar of the language. In the speech chain this is illustrated by transmitting the message through the linguistic channel, producing a word sequence W . Since a message can be formulated in many ways, the mapping is described by a probability, $P(W|M)$. The word sequence then triggers neuromuscular activity in the articulatory channel. This produces the spoken message, which is a sequence of sounds, S , radiating as an acoustic wave from the speaker's mouth. The articulatory channel will be different for every individual speaker; the shape and length of the vocal tract, the dimensions of the vocal folds and the muscles that control the articulators are different for each individual. This makes it impossible for two speakers to produce identical waveforms, indeed, it is near impossible for a speaker to reproduce a particular realization. The mapping of words to sounds is ruled by the probability $P(S|W)$.

The acoustic wave from the speaker's mouth will be further affected by the transmission channel from the mouth to the receiver to produce the received signal, X . The factors affecting the transmission channel will depend on the type of communication. In the simplest case of human speech communication this transmission channel will consist of the acoustic channel from the mouth to the ear, defined by the propagation properties of the room and the additive ambient noise. If the speech communication is conducted over the telephone, the channel will be combined of the acoustic channel from mouth to microphone (including effects of ambient noise on the transmitting end), the telephone transmission channel (including the transduction from acoustic to electric wave and back), and the acoustic channel on the receiving end. If we are looking at human-to-machine communication, the signal received by the speech recognizer will at least be affected by the acoustic channel from mouth to microphone and the characteristics of the microphone and A/D conversion. The mapping performed by the transmission channel is given by the probability $P(X|S)$.

At the receiving end, the human listener will try to reverse the production process to decode the original message. This process includes the analysis of the speech signal performed by the ear, which in turn produces electrical activity in the auditory nerve. The brain then applies knowledge of the language system in order to decode and comprehend the original message.

In order to better understand the rationales behind various approaches to speech recognition and synthesis, we will spend a little time looking at some fundamentals of speech production and perception before going into speech technology issues.

Speech Production

Speech is air pressure waves radiating from the mouth and nostrils of the speaker. The main components of the human speech production apparatus are the lungs, the glottis and the vocal tract. The driving source is air from the lungs. At the glottis, the vocal cords constrict the path from the lungs to the vocal tract. In voiced sounds, air pressure from the lungs build up behind the closed vocal cords, until they abruptly open to release a burst of air before closing again. The cycle repeats and produces a quasi-periodic sequence of excitation pulses. The inverse of the pulse period is called the fundamental frequency, which determines the perceived *pitch* of the speech signal. In unvoiced sounds, the vocal cords are open. The intonation of speech is determined by the variations of the fundamental frequency.

The excitation is filtered by the vocal tract to produce the sounds. The shape of the vocal tract, i.e. the position of the jaws, the opening of the lips, the shape of the tongue and the opening or closing of the velum will determine the frequency response of the vocal tract. Analysis of the vocal tract shows that the frequency response will typically be dominated by a small number of resonances, the *formants*. As we speak, we change the shape of the vocal tract, and thus the frequency response of the filter, in order to pro-

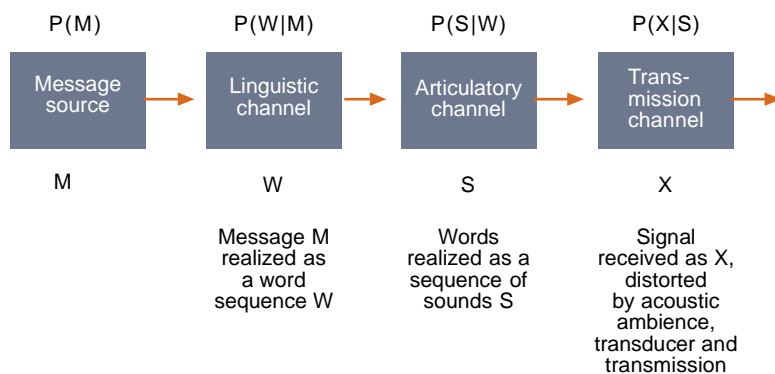


Figure 1 A communication-theoretic view of the speech communication chain (adapted from [16])

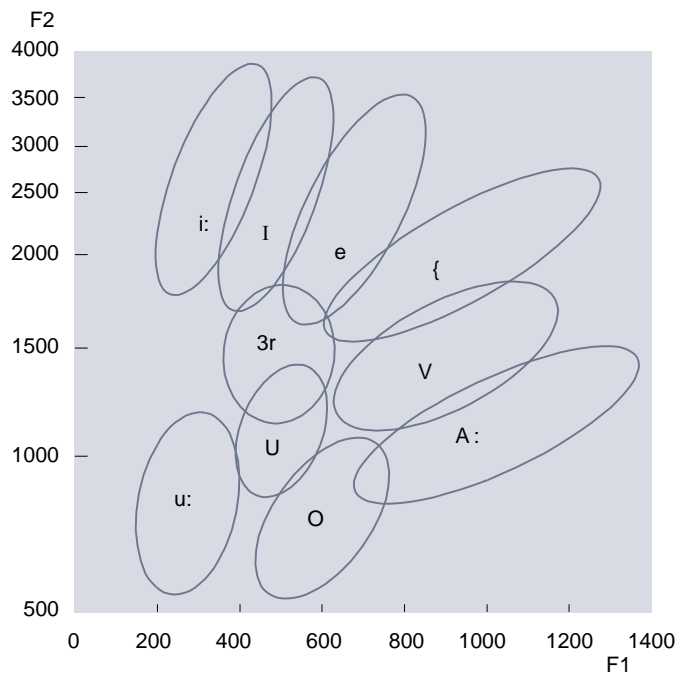


Figure 2 The vowel triangle showing the distribution of English vowels as a function of the first (F1) and second (F2) formants. The phoneme symbols are in SAMPA [35] notation

duce different sounds. The positions of the formants are the most significant factor in terms of human identification of speech sounds. Vowels can to a large extent be identified on the basis of the position of the two lowest formants, F1 and F2 (see Figure 2). However, the distribution of F1/F2 values for the different vowels has overlapping regions where the formant information is insufficient for making unambiguous decisions on vowel identity.

The smallest distinguishing unit of a language is the phoneme. The phoneme is an abstraction, covering a multitude of possible actual realizations, phones, of the sound. Different sounds are created by moving the articulators (e.g. tongue, lips, jaws). The articulators are physical entities with a mass, which means that their movement cannot be instantaneous. Thus, the realization of a phoneme will depend on the articulator positions of the preceding and the succeeding sounds. This phenomenon is called coarticulation. An accurate modeling of the coarticulation phenomena is vital for the performance of both speech recognizers and speech synthesizers.

Hearing and Perception

The human ear consists of three sections, the outer ear, the middle ear and the inner ear. The received air pressure wave travels through the outer ear and sets up a vibration at the eardrum. Vibration at the eardrum is picked up by the tiny bones of the middle ear which connect to the inner ear through the oval window of the cochlea. The cochlea is a liquid filled organ that connects directly to the auditory nerve. Vibrations at the oval window cause a displacement of the liquid along the basilar membrane of the

cochlea, the place of maximum displacement being dependent on the frequency content of the vibration. The displacement causes hair cells on the basilar membrane to be bent, resulting in neurological firings to the hearing nerve.

Human speech perception is dependent on the frequency of the stimulus. The cochlea performs a frequency analysis of the sound and decomposes it into spectral components, similar to a bank of overlapping bandpass filters where the bandwidth of each filter is increasing with the center frequency. It has been demonstrated that the bandwidth increase follows an exponential rule for frequencies above approximately 1 kHz and is close to linear below 1 kHz. This has prompted the use of the *mel* and *Bark* scales in speech analysis.

In terms of perception, the formants and the fundamental frequency are the factors that are the most important. Flanagan [13] reports results of a series of experiments aimed at determining the smallest deviations in isolated parameters for speech production that could be perceived by humans. The most sensitive parameter is the fundamental frequency where a deviation of 0.3–0.5 % could be detected. The frequencies and amplitudes of the two lowest formants were also parameters where small changes were easily detected. The formant bandwidths were less important, and humans seem to be fairly insensitive to deviations in the “spectral valleys”, between the formants.

Speech Technology

Speech technology comprises a number of different technologies. In the interest of space, this treatment will concentrate on automatic speech and speaker recognition and speech synthesis, excluding interesting areas such as speech coding, language identification and speech enhancement.

Speech Recognition

A speech recognizer is basically a device that upon the presentation of a speech utterance performs a predefined action. The predefined action can be to present a written transcription of the spoken utterance; to perform the action the utterance refers to (e.g. “Turn on the light”, “Open the last accessed document in Word”) or to extract the information in the spoken utterance that is relevant for a specific task and to interpret the meaning of that information (e.g.; if a user says “I need to go to the city to meet a friend, and I am running late. Can you tell me when the next bus from the University leaves?” the essential information a system for bus traffic information needs to extract is “Give departure time for next bus from University to city”).

A speech recognizer can be viewed as a classifier. The task of a classifier is to determine which class an observation belongs to. The number of classes is normally limited. For the speech recognizer the classes can be e.g. the words constituting the vocabulary of the recognizer. A typical measure of the quality of a classifier is the risk of misclassification. If we regard the speech recognizer as a word classifier and define the cost of erroneous classification uniformly to be 1 and the cost of correct recognition to 0, then the misclassification risk equals the word error rate, i.e. the expected percentage of words that will not be correctly recognized. The word error rate is the most widely used quality measure for speech recognizers.

The complexity of the task, and consequently the performance of the speech recognizer will depend on a number of factors. The most important are:

Vocabulary. The size of the vocabulary, i.e. the number of different words the recognizer is trained to recognize, will depend on the task the recognizer is designed for. Increasing the vocabulary size will increase the footprint of the recognizer, and it will also increase the risk of making errors. In addition, it is important to note that the content of the vocabulary can influence the performance of the system. Vocabularies containing many short, phonetically similar words are difficult to handle. This can be an important design factor for small vocabulary systems where it is possible to custom design the vocabulary content.

Speaking style. The speaking style can be isolated utterances or continuous speech. Isolated utterance systems are typically command and control applications, where the system only understands a limited number of commands. In continuous speech, the words of the vocabulary can be combined in any syntactically relevant manner to produce sentences. Continuous speech will typically allow an extremely high number of legal utterances, and will yield more complex systems than recognizers for isolated utterances. Large vocabulary continuous speech recognizers have traditionally required the user to speak in a carefully articulated manner, similar to reading. Spontaneous speech contains phenomena such as significant speaking rate variation, sloppy pronunciation, incomplete sentences and restarts, “uhms” and “ahs”, which most recognizers cannot handle. These artifacts are present e.g. in dictation, particularly if the user is not accustomed to dictating. Even more of these artifacts are present in conversations and dialogues, where problems like turn-taking and simultaneous speech are also present. The difficulties presented by speaking style is clearly illustrated when looking

at current state-of-the-art performance for conversational speech, which show word error rates around 35–40 %.

Speaker mode. The characteristics of speech depend on the speaker. This is partly due to physical differences in the vocal apparatus, e.g. differences in the length of the larynx, but dialects and accents will also contribute strongly to differences in the uttered speech. In order to achieve the best performance, a recognizer should ideally be trained on speech from the target speaker. This is impractical for most applications. Most modern speech recognizers are trained on speech from a large number of speakers, exposing the training algorithm to most normal variations in speech characteristics. For many applications this is sufficient to enable a satisfactory performance. In order to enhance the performance, the recognizer can adapt to the current speaker using whatever speech is available.

Speaker Recognition

In contrast to speech recognition, where the task is to identify *what* is spoken, the task of speaker recognition is to identify *who* is speaking. Speaker recognition can broadly be divided into two main application areas. The first is the task of verifying a speaker’s claimed or assumed identity, *speaker verification*, which is a binary (yes/no) classification problem. The second is *speaker identification*, which is to determine the identity of a speaker, either from a set of N known speakers (closed set), or from an unlimited set of speakers of which N are known to the system (open set). Closed set speaker identification is an N -class classification problem, while an open set speaker identifier can be viewed as an $(N+1)$ -class classifier where class $(N+1)$ represents all unknown speakers. As N increases, the probability of correct identification will decrease.

Speaker verification has received most interest, both because it is seen as having the highest potential for commercial applications (and is more reliable), and because many speaker identification applications can be recast to be implemented using speaker verification methods. Today, both speaker verification and speaker identification methods are based on very similar statistical paradigms. Thus, speaker verification will receive the greatest attention in this paper.

Speaker verification can be used for access control, which can include both physical access to restricted areas and logical access to various services. Examples of such services are telephone banking; telephone shopping; information access; travel services and computer account access. Speaker verification can also be applied to law enforcement by assessing the likelihood

that the voice in audio recordings made from e.g. telephone conversations is that of a specific person. It should however be noted that the use of speaker recognition is not admissible in a court of law in many countries. Another use of speaker recognition which lately has received considerable interest is speaker detection, or segmentation, which consists of identifying the current speaker in a multi-speaker audio recording such that segments of the recording can be automatically labeled with the identity of the speaker.

The intended application will determine the approaches to speaker recognition and to the reliability of the recognition. The main factors are:

Speech modality: Is the textual content of the speech known to the system or not? If the text is known, by requiring the user to speak a predefined or prompted phrase, system reliability can be very high. Such systems are termed text dependent. In text independent systems, the textual content is not known. The phonetic content of the input speech is thus also unknown, and the system cannot utilize knowledge about the speaker specific phone realizations, resulting in lower reliability.

Speech quality: The quality of the recording will have a strong impact on the performance of speaker recognition systems. Channel and ambient noise in an application where the user is calling the system from a mobile phone will cause degradation in the system performance relative to an application where the user is speaking directly into the microphone in a quiet environment. Similarly, forensic speaker recognition based on a (hidden) omnidirectional surveillance microphone in a reverberant room is far more difficult than recognition from a recording of a fixed line telephone conversation.

Amount of speech: Generally, a speaker recognition system will perform better the more speech can be made available for recognition. In many applications there will obviously be a trade-off between reliable identification and user convenience. A speaker recognition system requires speech data to train the speaker models, i.e. to “learn” the characteristics of the speakers. The amount of speech data available for training will have a strong influence on system performance.

Evaluation of system performance will be dependent on the boundary conditions defined by the application. For speaker verification systems there are two types of errors, *false acceptance*, i.e. that the identity claim/assumption is accepted although the speaker is an imposter, and *false rejection*, i.e. that the identity claim is rejected although the voice belongs to the claimed speaker. The severity of the two error

types will vary depending on the application, and deciding the relative costs is an important design criterion. False acceptance constitutes a security risk, while false rejections mainly impact on user convenience. A system designed for a high security application will thus attempt to minimize the false acceptances at the cost of accepting a higher rate of false rejections.

Speech Synthesis

Speech synthesis is generally the act of making the computer generate a spoken message from a textual concept. If the message vocabulary is small, the spoken message can be generated either by simple playback of pre-recorded messages or by *phrase concatenation*. During system design a library of pre-recorded phrases, i.e. parts of sentences, are recorded. For a talking clock a library consisting of the phrase “The time is now” and pronunciations of the numbers 0–20, 30, 40 and 50 would suffice to enable the computer to tell the time in hours and minutes. For example, at 21:37 hours, the clock can say the time by concatenating the carrier phrase “The time is now” with the words “twenty”; “one”; “thirty” and “seven”. In applications where the output messages are syntactically restricted and the number of messages is limited, a properly designed phrase concatenation system gives very high quality at a low cost. However, if it is desirable to change the message vocabulary a complete redesign is usually needed – unless the speaker that contributed the original recordings is available and the recording environment can be fairly closely reconstructed.

Although phrase concatenation has been widely used in applications with limited message vocabularies, the term “speech synthesis” is usually associated with text-to-speech synthesis (TTS). TTS implies that any text can be rendered as speech by the speech synthesizer. An important part of a TTS system will of course be the actual synthesizer, i.e. the conversion of a (textual) string of symbols to speech. But the TTS system must also perform a linguistic analysis of the input orthographic text in order to extract information about pronunciation, stress and timing, estimate where the emphasis should be put in a sentence, disambiguate words and phrases when multiple interpretations will yield different pronunciations, etc.

The quality of a TTS system will depend on both the linguistic processing and the speech synthesis. The basic linguistic processing will for instance ensure that numbers and abbreviations are correctly interpreted and pronounced and is thus an important factor in the overall quality assessment. However, the most widespread assessment strategies for TTS emphasize *naturalness* and *intelligibility*. Intelligibility is

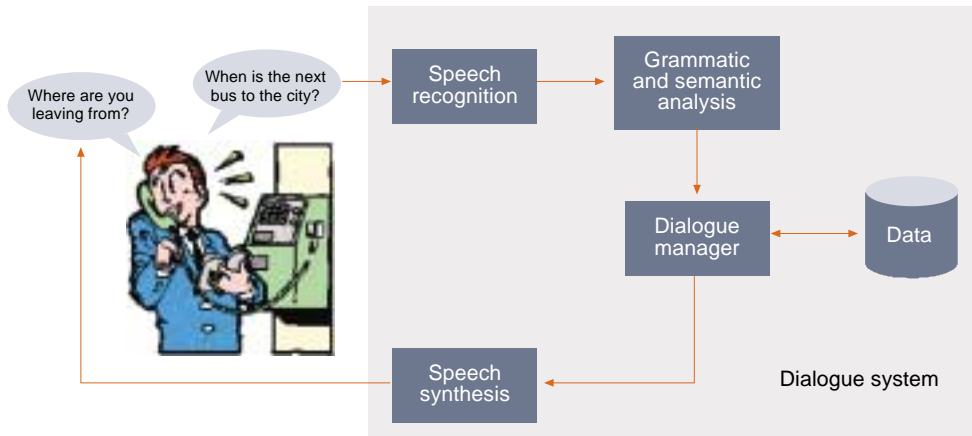


Figure 3 A spoken dialogue system for bus traffic information

often measured by the capability of human listeners to correctly identify the phonetic content of short, nonsense words or the words of semantically unpredictable sentences. Naturalness on the other hand is typically measured using techniques developed for evaluating speech compression algorithms, such as a subjective evaluation of quality by paired comparison or mean opinion score (MOS).

Spoken Dialogue Systems

In a spoken dialogue system, speech recognition is used to convert spoken requests and information to a symbolic form understandable by the information retrieval module. The speech recognizer can be used in conjunction with a semantic analysis module which attempts to extract the semantically meaningful part of the utterance, or the semantic processing can be integrated in the recognizer. Speech synthesis is employed to prompt the user for information and to convey the information the user has requested. In addition to the speech I/O modules, a vital part of a spoken dialogue system is the dialogue manager. The dialogue manager controls the dialogue, keeps tabs on the dialogue flow so that the system is not lost if the dialogue should not follow a linear path, performs recovery from errors made by the recognizer or the user, and acts as an interface between the information retrieval module and the speech modules. An example dialogue system is depicted in Figure 3.

The Past and Present

This section will attempt to give a brief overview of the most important developments in speech recognition and synthesis up till today. By necessity, the presentation will be rather superficial, leaving out many significant contributions to the development of speech technology.

Speech Synthesis

Early efforts at constructing speaking machines were undertaken as early as the second half of the 18th century. Wolfgang von Kempelen constructed a "speaking machine" [18] which was

an ingenious mechanical approximation of the human speech apparatus, using a bellows to simulate the function of the lungs, a reed to approximate the vocal cords and a flexible rubber tube could be manipulated to shape the "vocal tract" to produce the correct sound.

An electronic analogue to von Kempelen's machine was demonstrated by Homer Dudley at the 1939 World Fair [8]. In Dudley's VODER, the user could manipulate the resonances of the vocal tract, thereby affecting the sound characteristics, and the voicing and pitch could also be controlled by levers and pedals. Making the VODER speak was like playing an instrument, and although both the VODER and von Kempelen's speaking machine demonstrated that it was possible to make machines produce fairly intelligible speech, it could not be done automatically. The first attempt at automating the procedure was done in the early 1950s, when the pattern playback machine [6] was designed to produce a sound pattern that followed a spectrogram.

The early efforts at producing speaking machines focused more on the sound production than on text analysis. In fact, this has been the case up till today. Yet, the quality of the linguistic analysis of the text, including text normalization, homograph disambiguation, stress analysis and grapheme-to-phoneme conversion constitute will define the upper bound on the achievable quality of the complete TTS system. Still, here we will only remark that the text analysis is developing from a traditional, rule based approach to include more and more data driven approaches, e.g. by exploiting statistical methods for inferring general rules from exposure to examples. The remainder of this section will treat sound production.

From the early days there were two competing paradigms for speech synthesis. One approach aimed at modeling the articulators of the speech production mechanism in detail, *articulatory synthesis*. The other approach was targeted at

modeling the speech signal itself, through a source-filter model [10]. This is often also termed *terminal-analog* synthesis [12]. Modeling the signal is intrinsically simpler than modeling the articulators, and the terminal-analog synthesizers were dominant, with formant and linear predictive synthesizers being prominent examples. Articulatory synthesis still remains an active field of research.

The formant synthesizer is based on *synthesis-by-rule*. In a synthesis-by-rule synthesizer, rules for parameter values and trajectories corresponding to explicit phonetic phenomena are used to drive the synthesizer. The parameters are typically divided into source and filter parameters, e.g. voicing, durations, formant frequencies and bandwidths. The rules are extracted from parameterized speech data, and the extraction and refinement of the rules can be a grueling task. Perhaps the most prominent example of formant synthesizers is Dennis Klatt's Klattalk [19] which was later commercialized as DECTalk [5]. The strength of rule-based synthesis is that in principle, the rules can be applied to e.g. change the speaker characteristics and the speaking style without the necessity of a major redesign of the synthesizer. The clear disadvantage is that even in the best systems, the naturalness of the synthetic voice is limited.

The linear predictive synthesizer was the first example of *concatenative speech synthesis* which deviates from the synthesis-by-rule paradigm in that the synthesizer possesses a very limited explicit knowledge; most is embedded in the actual segments to be concatenated [9]. In a linear predictive synthesizer the filter parameters are generated from analyses of actual filter trajectories of speech segments. Since coarticulatory influences are smaller at the center of a phoneme, and the transitions between phonemes are perceptually crucial, the diphone was proposed as the basic synthesis unit. The diphone is a unit which stretches from the middle of one phone to the middle of the succeeding phone. The linear predictive model of speech is a simple source-filter model and lends itself readily to prosodic modification by altering the excitation (for pitch modification) or by interpolating the filter parameters (for modifying segment durations). Yet, the model has intrinsic weaknesses in terms of the achievable quality of the synthetic speech. The speech quality can be improved by introducing better production models, e.g. by the use of Multi-pulse excited LPC [1]. However, improving the production model comes at the cost of making prosodic control more difficult.

The next major step forward was the introduction of the *Pitch-Synchronous Overlap Add*

(PSOLA) technique [23]. The PSOLA approach allows manipulating the prosodic features of stored diphone waveforms, and the concatenation of the prosodically altered segments can be performed using the same paradigm. Thus, any desired sequence of phones exhibiting a desired prosody can in principle be generated from a limited diphone inventory. In practice there is a limit to what extent the prosody can be manipulated before it results in audible distortion. Also, the technique does not directly solve problems associated with spectral or phase discontinuities at the concatenation points. Nevertheless, the PSOLA technique was a quantum leap in terms of improving the naturalness of synthetic speech.

Diphone based PSOLA techniques depend on predicted prosodic characteristics, and the speech quality is vulnerable to discontinuities at the joints of the diphone segments. *Unit selection synthesis* (see e.g. [24, 4, 3]) is designed to avoid these disadvantages by eliminating, or greatly reducing, the need for prosodic manipulation of the stored waveforms. Instead of having a carefully designed diphone inventory with only a single, neutral realization of each diphone, unit selection synthesis is based on the availability of a large speech database, containing most of the natural prosodic variation. The basic idea is that if a desired phonetic and prosodic context can be found in a database of natural speech, nothing can be more natural sounding than that. The principle of unit selection synthesis is that the (appropriately processed and labeled) database is searched for a sequence of units that best matches the sequence that is predicted by the text analysis part of the TTS engine. The search of the database can be performed using dynamic programming. The cost of using a specific unit from the database is combined of two parts, the target cost, which is a measure of the match between the database unit and the desired unit, and the concatenation cost, which is a measure of the distortion that will arise by concatenating the unit with a preceding database unit. Letting the concatenation cost be zero for database units that are spoken in succession leads the search to favor unit sequences that occur naturally. This reduces the amount of spectral discontinuities and tends to imitate prosodic structures found in the database. Unit selection synthesis can produce very natural sounding speech, but can also fail miserably if the database does not contain good matches. Critics of the technique emphasize the data coverage problem as the main problem: due to the enormous variations in speech, it is impossible to collect (and use) a database that contains all possible phonetic and prosodic variation. However, allowing limited, high quality prosodic manipulation of the database units may reduce the validity of this criticism.

Speech Recognition

The first speech recognizer of sorts was demonstrated in the late 1920s. *Radio Rex* was a mechanical toy dog that would jump out of his house when his name was spoken. The action was basically triggered by the level of sound energy around 500 Hz, which is the frequency band around the first formant of the vowel 'e' (which implies that Rex would jump into action also as the result of other acoustic stimuli than his name). The first real speech recognizer was a device capable of recognizing the digits 0–9, developed by researchers at Bell Labs in the early 1950s [7]. This device identified the digits on the basis of a crude estimate of the formants of the vowels contained in the digits.

With the advent of the digital computer more sophisticated methods were made possible. In speech recognition the two most important contributions to the progress were the introduction of dynamic programming techniques for time-scale adjustments when comparing patterns and the introduction of robust and efficient methods for spectral estimation.

Temporal variation in speech tends to be non-linear. In order to compare two patterns of different length, but where the ratio of the durations of the speech sounds in the two patterns is not constant, the use of dynamic programming for time alignment was introduced ([17, 25]).

Dynamic Time Warping was a dominant technique for speech recognition until the mid-1980s, and systems using this simple method for matching time patterns can still be found. Still, most important was the introduction of dynamic programming methods to speech processing. Dynamic programming is a very powerful tool and is extensively used in current speech processing algorithms (see e.g. [21]).

The speech waveform itself exhibits too much variation to be well suited for speech recognition. Even in the early systems it was realized that the spectral domain is well suited for distinguishing between sounds. The introduction of short-term spectral estimation methods and corresponding distortion measures that were reasonably robust to speaker variations and which corresponded well with hearing constituted a major step forward. Several of these innovations came from outside the speech recognition community (e.g. linear prediction analysis, cepstral representation of the spectrum, distortion measures), but were readily applicable to ASR.

The early systems for speech recognition tended to adhere to two paradigms: pattern matching and/or acoustic-phonetics. Pattern matching systems based on Dynamic Time Warping tried to match incoming speech patterns with stored ref-

erence patterns, taking into account the non-linear time variation exhibited. The reference patterns were typically word-length, making it difficult to construct large vocabulary systems. Systems based on acoustic-phonetics originated in the study of the properties of the speech sounds, the *phonemes*. The phonemes can be classified depending on how and where in the vocal apparatus they are generated; i.e. the articulatory configuration typically observed for each sound. There are two main problems with the acoustic-phonetic approach: the difficulty of reliably estimating the articulatory features from a speech sample, and the insufficient description of the variability in real, spontaneous speech, often contaminated by noise. Thus, these approaches could be successful for limited vocabulary systems with carefully articulated speech, but were not adequate for solving the general problem of speech recognition.

A major step forward came with the introduction of statistical approaches for speech recognition. Hidden Markov Models (HMM) were proposed for speech recognition in the mid-70s ([2, 15]) and were within a decade to become the de facto standard approach to speech recognition. The success of the statistical approach, and HMM in particular, can be attributed to the following main factors:

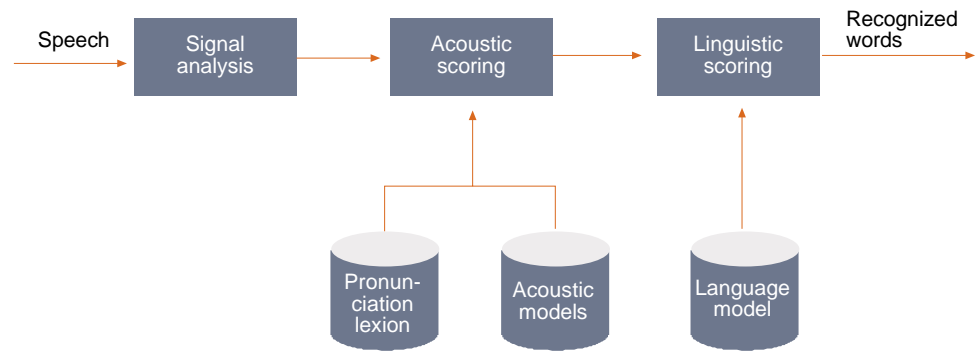
- 1 The statistical approach has the capability of modeling the variation found in normal speech.
- 2 The emergence of large databases made it possible to obtain reasonable parameter estimates for the complex statistical models needed to describe this variation.
- 3 The statistical framework is applicable to modeling the probabilistic mappings in the speech communication chain, in principle enabling a unified approach to decoding the underlying message.

Most research has concentrated on modeling the acoustic and linguistic mappings in Figure 1. The optimal classifier for determining the word sequence, $\mathbf{W} = (w_1, w_2, \dots, w_N)$ that is the source of an acoustic observation, \mathbf{X} , will select the sequence \mathbf{W}' that maximizes the a posteriori probability, i.e.

$$\begin{aligned} \mathbf{W}' &= \arg \max_{\mathbf{w}} \{P(\mathbf{W}|\mathbf{X})\} \\ &= \arg \max_{\mathbf{w}} \left\{ \frac{P(\mathbf{X}|\mathbf{W}) \cdot P(\mathbf{W})}{P(\mathbf{X})} \right\} \end{aligned} \quad (1)$$

Since the maximization is independent of $P(\mathbf{X})$, this term can be neglected. Thus, the choice of the optimal word sequence is dependent on two

Figure 4 Statistical speech recognizer



terms, the acoustic likelihood, $P(\mathbf{X}|\mathbf{W})$ and the language model $P(\mathbf{W})$. A simplified block diagram of a statistically based speech recognizer is shown in Figure 4.

The task of the statistical recognizer is thus to find the most likely word sequence for an observed spoken utterance, represented by a sequence of feature vectors produced by the signal analysis. In order to do this, statistical models for the various speech sounds must be estimated in a training phase. Normally, several models are produced for each speech sound, dependent on the phonetic context surrounding the target sound. This is done in order to model coarticulation effects. In order to describe the words of the recognizer vocabulary, a pronunciation lexicon defining the words in terms of phone sequences is required. Also, a language model must exist. This will in most cases be a statistical N-gram model, i.e. a model that estimates the probability of an N-tuple of words. For example, a trigram language model restricts the language modeling by assuming that the probability of a target word is only dependent on its two most recent predecessors, i.e. $P(w_n | w_{n-1} w_{n-2} \dots w_2 w_1) = P(w_n | w_{n-1} w_{n-2})$. Then, the probability of a word sequence (e.g. a sentence) $W = \{w_1, w_2, \dots, w_N\}$ can be found as

$$P(w_1 w_2 \dots w_N) \rightarrow P(w_1 w_2 \dots w_N) \quad (2)$$

Other notable advances in automatic recognition include the use of a perceptually based warping of the frequency axis when performing the signal analysis, and the use of a decorrelating (cosine) transform for producing mel-cepstrum feature vectors as the input to the statistical classifier. Mel-cepstrum feature vectors augmented by estimates of their first and second order time derivatives are standard features of current speech recognizers.

Speaker Recognition

The development of speaker recognition technology has to a large extent paralleled the evolution of speech recognition. As for a speech recognition system, an operational speaker recognition system will consist of an analysis

stage and a pattern matching stage. The analysis is aimed at extracting features that are robust and that are able to represent the speaker characteristics. Identity information that is embedded in the speech signal includes information on the physiology of the vocal tract, suprasegmental information like pitch and stress patterns and higher level information such as pronunciation, word and phrase frequencies and syntax. Suprasegmental and higher level information are relatively easy to mimic (these are features typically used by human mimics), and are not well suited for most applications of speaker recognition. The features used for speaker recognition are basically the same as used for speaker independent speech recognition, namely spectral features containing information on the vocal tract shape. This is intriguing, as the aim of the feature extraction in speaker recognition is to extract parameters that consistently and robustly emphasize the differences between speakers, while a speech recognizer front end should preserve the phonetic characteristics and aim to minimize the influence of speaker variations. However, both the main phonetic and the speaker dependent characteristics are dependent on the shape of the vocal tract, which is efficiently described by the smoothed spectral envelope that can be represented by e.g. cepstral or mel-cepstral parameters. Interestingly, the use of dynamic parameters (time derivatives of e.g. mel-cepstral parameters) was first proposed for speaker verification [27, 28] but is now common in both speech and speaker recognition systems.

The pattern matching stage compares the extracted features to stored models for each speaker known to the system in order to find the best match. Early systems used simple template matching techniques. More advanced template matching methods employing statistical feature averaging were introduced for text-independent speaker recognition in the 1970s. Here, a template is created for each speaker by averaging the feature vectors over the training speech. In recognition, a distance measure evaluating the deviation between the test speech and the stored template is evaluated and constitutes the basis for the decision. Dynamic Time Warping, which

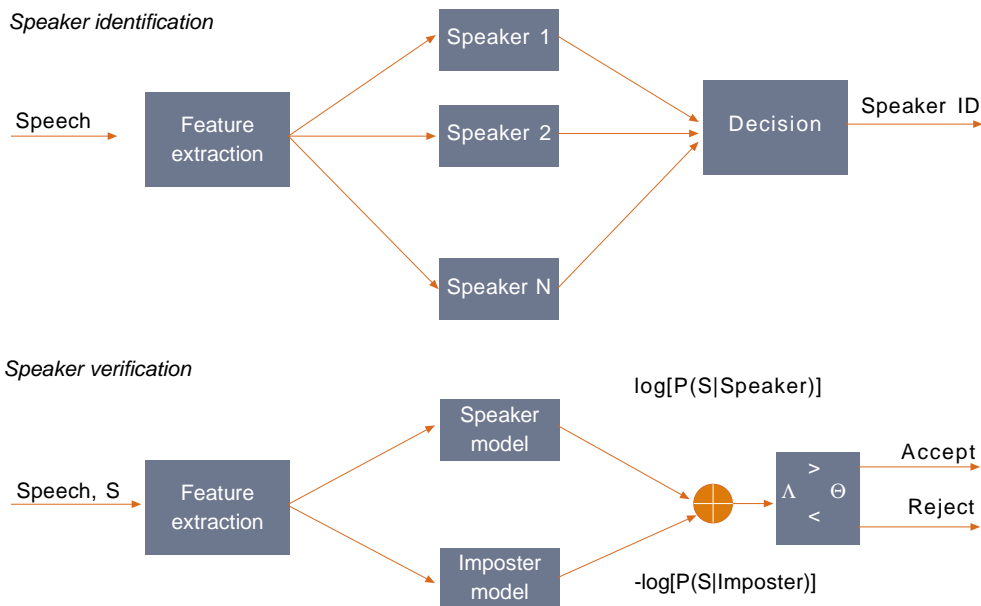


Figure 5 Speaker identification and verification

was introduced for speech recognition around 1970 was quickly adapted to speaker recognition, particularly for text-dependent applications. Here, templates representing words or utterances are created during training, and the test speech is compared and evaluated against these reference patterns.

The predominant approach to speaker recognition over the past 15 years has been based on probabilistic modeling. Nearest Neighbor modeling, which is a non-parametric approach, is based on representing each speaker's acoustic space by a collection of feature vectors. The collection can consist of all feature vectors in the training speech [29] or a vector quantization codebook designed from the training speech [30]. The distance between an input feature vector and its nearest neighbor is a measure of the probability of the assumed speaker, and is the basis for calculating the speaker score, which can be viewed as an estimate of the probability that the test speech emanates from a given speaker.

Yet, the parametric approach of Hidden Markov modeling has lately been the most successful approach to speaker recognition, as it has been for speech recognition. The basic HMM paradigm can be applied to both text-dependent and text-independent speaker recognition. In text-dependent speaker recognition, speaker models are created as word models or phone models, depending on the application. Using phone models, the speaker models can be used to verify any spoken word sequence. Word (or phrase) models restrict the application to a predefined vocabulary. For text-independent applications, a single-state HMM is used to model each speaker's acoustic space. This approach, usually referred

to as Gaussian Mixture Models (GMMs) [31], can also be viewed as modeling a speaker as a probabilistic source with a probability distribution that is a weighted sum of multivariate Gaussian densities. Recognition will then basically consist of estimating the likelihood of the speech being produced by a speaker and making a decision based on the estimated likelihoods.

The likelihood score obtained using HMM approaches can exhibit significant variation due to factors that are not related to the speaker, such as noise; microphone, channel and text variability. These effects can be reduced by processing the extracted feature by e.g. cepstral mean normalization [34]. An efficient method for reducing these effects are imposter models, which are used to normalize the likelihood score by letting the input speech be scored against both the target speaker and the imposter models [32, 33].

In speaker verification systems, it is vital to minimize the risk of accepting imposters. The algorithms for speaker verification take this into account by basing the decision on features that describe the vocal tract, a feature that is hard to mimic. However, in many applications one needs to consider the possibility of intruders using analog or digital recordings of a speaker. This can successfully break the system security if it is text-independent, or if fixed pass-phrases are used. Because of this, most current applications use prompted phrases. The system will generate a sentence which the user is asked to utter. In this way, the verification can be text-dependent (as the system knows which sentence to expect) but since the same sequence of words are never repeated; recordings cannot be used to defeat the system.

Brief Summary of the State-of-the-art

Speech synthesizers have during the last decade moved from being fairly intelligible but not very natural sounding to being (at best) quite natural sounding through the application of PSOLA-techniques and unit selection synthesis. For speech synthesis within a specific task domain, e.g. reading of news reports, unit selection synthesis currently produces (on average) quite natural sounding synthetic speech, although the prosodic variation is more limited than in human speech. For general reading machines, where it is desirable to produce a multitude of speaking styles, including the ability to express emotions, unit selection synthesis is currently not applicable. It is still an open question whether this will be attainable using corpus based speech synthesis.

Two of the major problems associated with unit selection synthesis is the footprint of the system and the complexity of creating new voices. The size of the database used in the synthesis will reflect on both memory requirement of the synthesizer and the computational requirements. Database design and organization are important issues if unit selection synthesis shall be available e.g. for handheld devices. Creating a new voice will include the recording and annotation of a large speech database. In particular, the processing and annotation of the data is time consuming and costly if done manually. Recent efforts at automating the process [26] have shown very promising results, and can lead to great savings in time and cost of creating new synthetic voices.

The state-of-the-art in speech recognition is mainly dependent on speaking style and the noise environment. For clean speech, i.e. when the effects of ambient or transmission noise is negligible, the performance of current speech recognizers is adequate for a large number of tasks as long as the input speech is fairly well articulated and fluent. Dictation systems for limited domains, such as radiology, where the amount of variation in syntax and vocabulary is limited, and the speakers are used to dictation, have been well received. Good performance for general speech-to-text remains a more elusive goal. When it comes to spontaneous and conversational speech, the performance of current speech recognizers is far below what is acceptable for most applications. Systems for automatically transcribing spontaneous monologues or natural conversations are beyond the capabilities of today's technology.

One of the main problems for current speech recognizers is the performance in the presence of noise. Humans have little problems with understanding noisy speech, even at high noise

levels, but speech technology has yet to find good methods for handling the effects of additive or convolutional noise. Many approaches have been investigated, but the problem is far from solved. Even relatively simple tasks, such as isolated word recognition for command and control applications, can be severely affected if the noise conditions do not match the environment for which the recognizer was trained.

Systems for spoken dialog, e.g. for automated information retrieval, have been successfully deployed. The complexity and performance of these systems depend on the dialog structure adopted. System driven dialog systems for restricted tasks can be well performing, particularly if techniques such as word spotting or concept spotting is employed, allowing words not relevant for the task to be ignored or even mis-recognized. Mixed initiative, or even user initiative systems such as AT&T's "How may I help you?" system for routing telephone calls [14], are not quite as readily deployed, although very encouraging results are obtained.

The Future

One of the major trends that can be observed from the progress in speech technology over the past decades is that there is a convergence of the techniques employed in the various disciplines. It is a clear tendency to rely on data driven methods, using statistical tools for modeling. In speech synthesis this is exemplified by the corpus based unit selection synthesis, the HMM-based methods for segmenting and labeling the speech databases, and also statistical methods for linguistic analysis.

In speech recognition, statistics has ruled the ground for the past 15 years, dominating both the acoustic modeling and the language and grammar. Even for semantic processing, statistical methods have shown great promise. On the other hand, speech recognition has over the same period of time moved from being a more or less pure signal processing discipline to requiring the use of explicit linguistic knowledge. This is of course connected to the evolution of the task complexity, which in many cases today requires the application of knowledge of linguistic, phonetic and discourse structures.

One interesting recent development is the interest in finite state automata for speech recognition. Although finite state networks have been a useful representation of both simple grammars as well as the state networks used for HMM decoding, work at AT&T has introduced a weighted finite state transducer (FST) formalism for describing the entire decoding process, including the acoustic models, the vocabulary and pronunciation lexicon, context dependency

and language model [22]. The immediate advantage of the FST formulation is that it allows for designing faster and more efficient decoders. However, the formalism can also give new insights into speech recognition and allows for a very interesting connection to the use of finite state machine formalisms in text-to-speech synthesis and in computational linguistics.

In spite of the convergence of techniques, there is still no “Grand unifying theory of speech” that can guide the way to better speech technology. We still lack some of the fundamental understanding of speech and language communication that can produce truly speech *understanding* systems, generation of natural sounding (emotional) synthetic speech from a concept, and *conversational* machines. We are currently making up for some of the lack of theory by employing statistics, with some degree of success.

Conclusions

Speech technology has come a long way from its primitive origins, and the technology is currently at a stage where a significant number of useful and profitable applications can be made. Still, it is important to be aware of the limitations of the current technology, both for designing good and user-friendly systems today, and for aiming to solve the challenges of tomorrow. Speech technology is not mono-disciplinary, and in order to be able to create a speech-enabled machine that can pass the Turing test, the collaborative efforts of scientists from many areas are needed.

Among the immediate challenges are: how to design speech recognizers that are robust to speaker accents and dialects as well as to noise contamination, how to create speech synthesizers that are able to convey emotions and can adopt speaking styles appropriate for any given text, how to deal with multi-linguality, how to attack the problem of recognizing spontaneous speech, just to name a few. Many problems have been solved, but speech technology research will not lack challenges in the years to come.

References

- 1 Atal, B S, Remde, J R. A new model for LPC excitation for producing natural-sounding speech at low bit rates. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Paris, 614–617, 1982.
- 2 Baker, J K. The DRAGON system – An overview. *IEEE Trans. Acoustics, Speech and Signal Processing*, 23 (1), 24–29, 1975.
- 3 Beutnagel, M et al. *The AT&T NextGen TTS system*. Joint meeting of ASA, EAA and DAGA, Berlin, 1999.
- 4 Black, A, Campbell, N. Optimising selection of units from speech databases for concatenative synthesis. *Proc. Eurospeech '95*, Madrid, 581–584, 1995.
- 5 Bruckert, E, Minow, M, Tetschner, W. Three-Tiered Software and VLSI Aid Developmental System to Read Text Aloud. *Electronics*, 21 April 1983.
- 6 Cooper, F S et al. Some experiments on the perception of synthetic speech sounds. *J. Acoust. Soc. Am.*, 24, 739–764, 1952.
- 7 Davis, K H, Biddulph, R, Balashek, S. Automatic Recognition of Spoken Digits. *Journal of the Acoustical Society of America*, 24 (6), 637–64, 1952.
- 8 Dudley, H W, Riesz, R R, Watkins, S A. A Synthetic Speaker. *Journal of the Franklin Institute*, 227, 739–764, 1939.
- 9 Dutoit, T. *An Introduction to Speech Synthesis*. Dordrecht, Kluwer Academic Publishers, 1997.
- 10 Fant, G. *Acoustic theory of speech production*. The Hague, Mouton and co., 1960.
- 11 Ferguson, J. *Hidden Markov Models for Speech*. Princeton, NJ, IDA, 1980.
- 12 Flanagan, J L. Notes on the design of Terminal-Analog speech synthesizers. *J. Acoust. Soc. Am.*, 29, 306–310, 1957.
- 13 Flanagan, J L. *Speech analysis, synthesis and perception, 2nd Edition*. Berlin, Springer-Verlag, 1972.
- 14 Gorin, A, Riccardi, G, Wright, J H. How May I Help You. *Speech Communication*, 23, 113–127, 1997.
- 15 Jelinek, F, Bahl, L R, Mercer, R L. Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Trans. Information Theory*, IT-21, 250–256, 1975.
- 16 Juang, B-H, Furui, S. Automatic recognition and understanding of spoken language – a first step toward natural human-machine communication. *Proc. IEEE*, 88 (8), 1142–1165, 2000.
- 17 Viyntuk, T K. Speech discrimination by dynamic programming. *Kibernetika*, 4, 81–88, Jan.-Feb. 1968.

- 18 von Kempelen, W. *Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine*, 1791.
- 19 Klatt, D H. The Klattalk text-to-speech system. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Paris, 1589–1592, 1982.
- 20 Klatt, D H. Review of text-to-speech conversion for English. *J. Acoust. Soc. Am.*, 92 (6), 737–793, 1987.
- 21 Lee, C-H. Applications of dynamic programming to speech and language processing. *AT&T Technical Journal*, 68 (3), 114–130, 1989.
- 22 Mohri, M, Pereira, F, Riley, M. Weighted finite-state transducers in speech recognition. *Computer Speech and Language*, 16, 69–88, 2002.
- 23 Moulines, E, Charpentier, F. Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9 (5-6), 453–467, 1990.
- 24 Sagisaka, Y et al. ATR – v-TALK speech synthesis system. *Proc. ICSLP'92*, Banff, 483–486, 1992.
- 25 Sakoe, H, Chiba, S. A Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Trans. on ASSP*, 26 (27), 43–49, 1978.
- 26 Syrdal, A et al. Corpus-based techniques in the AT&T NextGen synthesis system. *Proc. ICSLP 2000*, Beijing, 2000.
- 27 Sagayama, S, Itakura, F. On Individuality in a Dynamic Measure of Speech. *Proc. ASJ Spring Conf. 1979*, 3-2-7, 589–590, June 1979.
- 28 Furui, S. Cepstral Analysis Technique for Automatic Speaker Verification. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 29, 254–272, 1981.
- 29 Higgins, A L, Bahler, L G, Porter, J E. Voice identification using nearest-neighbor distance measure. *Proc. ICASSP'93*, Minneapolis, II, 375–378, 1993
- 30 Soong, F K et al. A vector quantization approach to speaker recognition. *Proc. ICASSP'85*, Tokyo, 387–390, 1985.
- 31 Reynolds, D A, Rose, R C. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. *IEEE Trans. on Speech and Audio Processing*, 3, 72–83, 1995.
- 32 Higgins, A L, Bahler, L G, Porter, J E. Speaker verification using randomized phrase prompting. *Digital Signal Processing*, 1, 89–106, 1991.
- 33 Carey, M, Parris, E, Bridle, J. A speaker verification system using Alphanets. *Proc. ICASSP'91*, 397–400, 1991.
- 34 Atal, B. Automatic recognition of speakers from their voices. *Proceedings of the IEEE*, 64, 460–475, April 1976.
- 35 *SAMPA Computer Readable Phonetic Alphabet*. July 29, 2003 [online] – URL: <http://www.phon.ucl.ac.uk/home/sampa>