50 years of progress in speech and speaker recognition

Sadaoki Furui

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

Abstract

Research in automatic speech and speaker recognition has now spanned five decades. This paper surveys the major themes and advances made in the past fifty years of research so as to provide a technological perspective and an appreciation of the fundamental progress that has been accomplished in this important area of speech communication. Although many techniques have been developed, many challenges have yet to be overcome before we can achieve the ultimate goal of creating machines that can communicate naturally with people. Such a machine needs to be able to deliver a satisfactory performance under a broad range of operating conditions. A much greater understanding of the human speech process is required before automatic speech and speaker recognition systems can approach human performance.

1. Introduction

Speech is the primary means of communication between humans. For reasons ranging from technological curiosity about the mechanisms for mechanical realization of human speech capabilities to the desire to automate simple tasks which necessitate human-machine interactions, research in automatic speech and speaker recognition by machines has attracted a great deal of attention for five decades.

Based on major advances in statistical modeling of speech, automatic speech recognition systems today find widespread application in tasks that require human-machine interface, such as automatic call processing in telephone networks and query-based information systems that provide updated travel information, stock price quotations, weather reports, etc.

This paper reviews major highlights during the last five decades in the research and development of automatic speech and speaker recognition so as to provide a technological perspective. Although many technological progresses have been made, there still remain many research issues that need to be tackled.

2. Speech recognition

The progress of automatic speech recognition (ASR) technology in the past 50 years can be summarized as follows [63, 33, 24]:

2.1. 1950s and 1960s

(1) ***General***: The earliest attempts to devise ASR systems were made in 1950s and 1960s, when various researchers tried to exploit fundamental ideas of acoustic phonetics. Since signal processing and computer technologies were yet very primitive, most of the speech recognition systems investigated used spectral resonances during the vowel region of each utterance which were extracted from output signals of an analogue filter bank and logic circuits.

(2) ***Early systems***: In 1952, at Bell Laboratories, Davis, Biddulph, and Balashek built a system for isolated digit recognition for a single speaker [11], using the formant frequencies measured/estimated during vowel regions of each digit. In an independent effort at RCA Laboratories in 1956, Olson and Belar tried to recognize 10 distinct syllables of a single speaker, as embodied in 10 monosyllabic words [57]. In 1959, at University College in England, Fry and Denes tried to build a phoneme recognizer to recognize four vowels and nine consonants [17]. By incorporating statistical information concerning allowable phoneme sequences in English, they increased the overall phoneme recognition accuracy for words consisting of two or more phonemes. This work marked the first use of statistical syntax (at the phoneme level) in automatic speech recognition. In 1959, Forgie and Forgie at MIT Lincoln Laboratories devised a system which was able to recognize 10 vowels embedded in a /b/ - vowel - /t/ format in a speaker-independent manner [16]. In the 1960s, since computers were still not fast enough, several special-purpose hardwares were built. Suzuki and Nakata at the Radio Research Lab in Japan built a hardware vowel recognizer [79]. Sakai and Doshita at Kyoto University built a hardware phoneme recognizer in 1962, using a hardware speech segmenter and a zero-crossing analysis of different regions of the input utterance [70]. Nagata and his colleagues at NEC Laboratories built a hardware digit recognizer in 1963 [55].

(3) ***DTW***: One of the difficult problems of speech recognition exists in the nonuniformity of time scales in speech events. In the 1960s, Martin and his colleagues at RCA Laboratories developed a set of elementary time-normalization methods, based on the ability to reliably detect speech starts and ends, that significantly reduced the variability of the recognition scores [49]. Martin ultimately founded one of the first speech recognition companies, Threshold Technology. At about the same time, in the Soviet Union, Vintsyuk proposed the use of dynamic programming methods for time aligning a pair of speech utterances (generally known as dynamic time warping (DTW)), including algorithms for connected word recognition [84]. However, his work was largely unknown in other countries until the 1980s. At the same time, in an independent effort in Japan, Sakoe and Chiba at NEC Laboratories also started to use a dynamic programming technique to solve the nonuniformity problem [72]. Since the late 1970s, dynamic programming in numerous variant forms, including the Viterbi algorithm [85] which came from the communication theory community, has become an indispensable technique in automatic speech recognition.

(4) ***Continuous speech recognition***: In the late 1960s, Reddy at Carnegie Mellon University conducted a pioneering research in the field of continuous speech recognition by dynamic tracking of phonemes [65].

2.2. 1970s

(1) ***General***: In the 1970s, speech recognition research achieved a number of significant mile stones. First, the area of isolated word or discrete utterance recognition became a viable and usable technology based on fundamental studies in Russia and Japan. Velichko and Zagoruyko in Russia advanced the use of pattern-recognition ideas in speech recognition [83]. Sakoe and Chiba advanced their techniques of using dynamic programming; and Itakura, when he was staying at Bell laboratories, showed how the ideas of linear predictive coding (LPC) could be extended to speech recognition systems through the use of an appropriate distance measure based on LPC spectral parameters [29].

(2) ***IBM Labs***: Researchers started studying large vocabulary speech recognition for three distinct tasks, namely the New Raleigh language for simple database queries [80], the laser patent text language for transcribing laser patents [30], and the office correspondence task, called Tangora, for dictation of simple memos.

(3) ***AT&T Bell Labs***: Researchers began a series of experiments aimed at making speaker-independent speech-recognition systems [64]. To achieve this goal, a wide range of sophisticated clustering algorithms were used to determine the number of distinct patterns required to represent all variations of different words across a wide user population.

(4) ***DARPA program***: An ambitious speech understanding project was funded by the Defense Advanced Research Projects Agency (DARPA), which led to many seminal systems and technologies [37]. One of the first demonstrations of speech understanding was achieved by CMU in 1973. Their Hearsay I system was able to use semantic information to significantly reduce the number of alternatives considered by the recognizer. CMU's Harpy system [48] was shown to be able to recognize speech using a vocabulary of 1,011 words with reasonable accuracy. One particular contribution from the Harpy system was the concept of graph search, where the speech recognition language is represented as a connected network derived from lexical representations of words, with syntactical production rules and word boundary rules. The Harpy system was the first to take advantage of a finite state network (FSN) to reduce computation and efficiently determine the closest matching string.

Other systems developed under the DARPA's speech understanding program included CMU'S Hearsay II and BBN'S HWIM (Hear What I Mean) systems [37]. The approach proposed by Hearsay II of using parallel asynchronous processes that simulate the component knowledge sources in a speech system was a pioneering concept. A global "blackboard" was used to integrate knowledge from parallel sources to produce the next level of hypothesis.

2.3. 1980s

(1) ***General***: The problem of creating a robust system capable of recognizing a fluently spoken string of connected word (e.g., digits) was a focus of research in the 1980s. A wide variety of the algorithms based on matching a concatenated pattern of individual words were formulated and implemented, including the two-level dynamic programming approach by Sakoe at NEC [71], the one-pass method by Bridle and Brown at Joint Speech Research Unit (JSRU) in UK [8], the level-building approach by Myers and Rabiner at Bell Labs [54], and the frame-synchronous level-building approach by Lee and Rabiner at Bell Labs [39]. Each of these "optimal" matching procedures had its own implementation advantages.

(2) ***Statistical modeling***: Speech recognition research in the 1980s was characterized by a shift in methodology from the more intuitive template-based approach (a straightforward pattern recognition paradigm) towards a more rigorous statistical modeling framework. Today, most practical speech recognition systems are based on the statistical framework developed in the 1980s and their results, with significant additional improvements having been made in the 1990s.

(3) ***HMM***: One of the key technologies developed in the 1980s is the hidden Markov model (HMM) approach [15, 62, 63]. It is a doubly stochastic process in that it has an underlying stochastic process that is not observable (hence the term hidden), but can be observed through another stochastic process that produces a sequence of observations. Although the HMM was well known and understood in a few laboratories (primarily IBM, Institute for Defense Analysis (IDA) and Dragon Systems), it was not until widespread publication of the methods and theory of HMMs in the mid-1980s that the technique became widely applied in virtually every speech recognition research laboratory in the world.

(4) ***Δcepstrum***: Furui proposed to use the combination of instantaneous cepstral coefficients and their first and second-order polynomial coefficients, now called Δ and ΔΔcepstral coefficients, as fundamental spectral features for speech recognition [21]. He proposed this method for speaker recognition in the late 1970s, but no one attempted to apply it to speech recognition for many years. This method is now widely used in almost all speech recognition systems.

(5) ***N-gram***: A primary focus of IBM was the development of a structure of a language model (grammar), which was represented by statistical syntactical rules describing how likely, in a probabilistic sense, was a sequence of language symbols (e. g., phonemes or words) that could appear in the speech signal. The *n-gram* model, which defined the probability of occurrence of an ordered sequence of *n* words, was introduced, and, since then, the use of *n-gram* language models, and its variants, has become indispensable in large-vocabulary speech recognition systems [31].

(6) ***Neural net***: In the 1980s, the idea of applying neural networks to speech recognition was reintroduced. Neural networks were first introduced in the 1950s, but they did not prove useful because of practical problems. In the 1980s, a deeper understanding of the strengths and limitations of the technology was achieved, as well as an understanding of the relationship of this technology to classical pattern classification methods [35, 45, 86].

(7) ***DARPA program***: The DARPA community conducted research on large-vocabulary, continuous-speech recognition systems, aiming at achieving high word accuracy for a 1000-word database management task. Major research contributions resulted from efforts at CMU with the SPHINX system [41, BBN with the BYBLOS system [10], SRI with the DECIPHER system [87], Lincoln Labs [58], MIT [89] and AT&T Bell Labs [40]. The SPHYNX system successfully integrated the statistical method of HMM with the network search strength of the earlier Harpy system. Hence, it was able to train and embed context-dependent phone models in a sophisticated lexical decoding network.

2.4. 1990s

(1) ***General***: In the 1990s, a number of innovations took place in the field of pattern recognition. The problem of pattern recognition, which traditionally followed the framework of Bayes and required estimation of distributions for the data, was transformed into an optimization problem involving minimization of the empirical recognition error [32]. This fundamental paradigmatic change was caused by the recognition of the fact that the distribution functions for the speech signal could not be accurately chosen or defined, and that Bayes' decision theory becomes inapplicable under these circumstances. Fundamentally, the objective of a recognizer design should be to achieve the least recognition error rather than provide the best fitting of a distribution function to the given (known) data set as advocated by the Bayes criterion. This error minimization concept produced a number of techniques, such as discriminative training and kernel-based methods.

As an example of discriminative training, the Minimum Classification Error (MCE) criterion was proposed along with a corresponding Generalized Probabilistic Descent (GPD) training algorithm to minimize an objective function which acts to approximate the error rate closely [9]. Another example was the Maximum Mutual Information (MMI) criterion. In MMI training, the mutual information between the acoustic observation and its correct lexical symbol averaged over a training set is maximized. Although this criterion is not based on a direct minimization of the classification error rate and is quite different from the MCE based approach, it is well founded in information theory and possesses good theoretical properties. Both the MMI and MCE can lead to speech recognition performance superior to the maximum likelihood based approach [9].

(2) ***DARPA program***: The DARPA program continued into the 1990s, with emphasis shifting to natural language front ends to the recognizer. The central focus also shifted to the task of retrieving air travel information, the Air Travel Information Service (ATIS) task. Later the emphasis was expanded to a range of different speech-understanding applications areas, in conjunction with a new focus on transcription of broadcast news (BN) and conversational speech. The Switchboard task is among the most challenging ones proposed by DARPA; in this task speech is conversational and spontaneous, with many instances of so-called disfluencies such as partial words, hesitation and repairs. The BN transcription technology was integrated with information extraction and retrieval technology, and many application systems, such as automatic voice document indexing and retrieval systems, were developed. A number of human language technology projects funded by DARPA in the 1980s and 1990s further accelerated the progress, as evidenced by many papers published in *The Proceedings of the DARPA Speech and Natural Language/Human Language Workshop*.

(3) ***Robust speech recognition***: Various techniques were investigated to increase the robustness of speech recognition systems against the mismatch between training and testing conditions, caused by background noises, voice individuality, microphones, transmission channels, room reverberation, etc. Major techniques include the maximum likelihood linear regression (MLLR) [42], the model decomposition [82], parallel model composition (PMC) [26], and the structural maximum a posteriori (SMAP) method [74].

(4) ***Applications***: Speech recognition technology was increasingly used within telephone networks to automate as well as enhance operator services.

2.5. 2000s

(1) ***DARPA program***: The Effective Affordable Reusable Speech-to-Text (EARS) program was conducted to develop speech-to-text (automatic transcription) technology with the aim of achieving substantially richer and much more accurate output than before. The tasks include detection of sentence boundaries, fillers, and disfluencies. The program was focusing on natural, unconstrained human-human speech from broadcasts and foreign conversational speech in multiple languages. The goal was to make it possible for machines to do a much better job of detecting, extracting, summarizing, and translating important information, thus enabling humans to understand what was said by reading transcriptions instead of listening to audio signals [47, 76].

(2) ***Spontaneous speech recognition***: Although read speech and similar types of speech, e.g. news broadcasts reading a text, can be recognized with accuracy higher than 95% using state-of-the-art speech recognition technology, recognition accuracy drastically decreases for spontaneous speech. Broadening the application of speech recognition depends crucially on raising recognition performance for spontaneous speech. In order to increase recognition performance for spontaneous speech, several projects have been conducted. In Japan, a 5-year national project "Spontaneous Speech: Corpus and Processing Technology" was conducted. A world-largest spontaneous speech corpus, "Corpus of Spontaneous Japanese (CSJ)" consisting of approximately 7 millions of words, corresponding to 700 hours of speech, was built, and various new techniques were investigated. These new techniques include flexible acoustic modeling, sentence boundary detection, pronunciation modeling, acoustic as well as language model adaptation, and automatic speech summarization [23, 25].

(3) ***Robust speech recognition***: To further increase the robustness of speech recognition systems, especially for spontaneous speech, utterance verification and confidence measures are being intensively investigated [38]. In order to have intelligent or human-like interactions in dialogue applications, it is important to attach to each recognized event a number that indicates how confidently the ASR system can accept the recognized events. The confidence measure serves as a reference guide for a dialogue system to provide an

appropriate response to its users. To detect semantically significant parts and reject irrelevant portions in spontaneous utterances, a detection-based approach has recently been investigated [36]. This combined recognition and verification strategy works well especially for ill-formed utterances.

In order to build acoustic models more sophisticated than conventional HMMs, the dynamic Bayesian network has recently been investigated [90].

(4) ***Multimodal speech recognition***: Humans use multimodal communication when they speak to each other. Studies in speech intelligibility have shown that having both visual and audio information increases the rate of successful transfer of information, especially when the message is complex or when communication takes place in a noisy environment. The use of the visual face information, particularly lip information, in speech recognition has been investigated, and results show that using both types of information gives better recognition performances than using only the audio or only the visual information, particularly in noisy environment.

3. Speaker recognition

Topics of the progress of automatic speaker recognition technology in the past 50 years can be summarized as follows:

3.1. 1960s and 1970s

(1) ***Early systems***: The first attempts for automatic speaker recognition were made in the 1960s, one decade later than that for automatic speech recognition. Pruzansky at Bell Labs [60] was among the first to initiate research by using filter banks and correlating two digital spectrograms for a similarity measure. Pruzansky and Mathews [61] improved upon this technique; and, Li et al. [44] further developed it by using linear discriminators. Doddington at Texas Instruments (TI) [12] replaced filter banks by formant analysis.

Intra-speaker variability of features, one of the most serious problems in speaker recognition, was intensively investigated by Endres et al. [14] and Furui [18].

(2) ***Text-independent methods***: For the purpose of extracting speaker features independent of the phonetic context, various parameters were extracted by averaging over a long enough duration or by extracting statistical or predictive parameters. They include averaged auto-correlation [7], instantaneous spectra covariance matrix [43], spectrum and fundamental frequency histograms [4], linear prediction coefficients [73], and long-term averaged spectra [19].

(3) ***Text-dependent methods***: Since the performance of text-independent methods was limited, time-domain and text-dependent methods were also investigated [2, 3, 20, 68]. In time-domain methods, with adequate time alignment, one can make precise and reliable comparisons between two utterances of the same text, in similar phonetic environments. Hence, text-dependent methods have a much higher level of performance than text-independent methods.

(4) ***Texas Instruments system***: TI built the first fully automated large scale speaker verification system providing high operational security. Verification was based on a four-word randomized utterance built from a set of 16 monosyllabic words. Digital filter banks were used for spectral analysis, and the decision strategy was sequential requiring up to 4 utterances for the trial. Several millions of tests were made over a period of 6 years for several hundred of speakers.

(5) ***Bell Labs system***: The Bell Labs built experimental systems aimed to work over dialed-up telephone lines. Furui [20] proposed using the combination of cepstral coefficients and their first and second polynomial coefficients as frame-based features to increase robustness against distortions by the telephone system. He implemented an online system and tested it for a half year with many calls by 120 users. The cepstrum-based features later became standard, not only for speaker recognition, but also for speech recognition.

3.2. 1980s

(1) ***HMM-based text-dependent methods***: As an alternative to the template-matching approach for text-dependent speaker recognition, the HMM technique was introduced in the same way for speech recognition. HMMs have the same advantages for speaker recognition as they do for speech recognition. Remarkably robust models of speech events can be obtained with only small amounts of specification or information accompanying training utterances. Speaker recognition systems based on an HMM architecture used speaker models derived from a multi-word sentence, a single word, or a phoneme. Typically, multi-word phrases (a string of seven to ten digits, for example) were used, and models for each individual word and for “silence” were combined at a sentence level according to a predefined sentence-level grammar [56].

(2) ***VQ/HMM-based text-independent methods***: Nonparametric and parametric probability models were investigated for text-independent speaker recognition. As a nonparametric model, vector quantization (VQ) was investigated [77, 69]. A set of short-time training feature vectors of a speaker can be efficiently compressed to a small set of representative points, a so-called VQ codebook. A matrix quantizer encoding multi-frame was also investigated [78, 34]. As a parametric model, HMM was investigated. Pritz [59] proposed using an ergodic HMM (i.e., all possible transitions between states are allowed). An utterance was characterized as a sequence of transitions through a 5-state HMM in the acoustic feature space. Tishby [81] expanded Poritz’s idea by using an 8-state ergodic autoregressive HMM represented by continuous probability density functions with 2 to 8 mixture components per state, which had a higher spectral resolution than the Poritz’s model. Rose et al. [67] proposed using a single-state HMM, which is now called Gaussian mixture model (GMM), as a robust parametric model.

3.3. 1990s

(1) ***Robust recognition***: Research on increasing robustness became a central theme in the 1990s. Matsui et al. [50] compared the VQ-based method with the discrete/continuous ergodic HMM-based method, particularly from the viewpoint of robustness against utterance variations. They found that the continuous ergodic HMM method is far superior to the discrete ergodic HMM method and that the continuous ergodic HMM method is as robust as the VQ-based method when enough training data is available. They investigated speaker identification rates using the continuous HMM as a function of the number of states and mixtures. It was shown that speaker recognition rates were strongly correlated with the total

number of mixtures, irrespective of the number of states. This means that using information about transitions between different states is ineffective for text-independent speaker recognition and, therefore, GMM achieves almost the same performance as the multiple-state ergodic HMM.

(2) ***Text-prompted method***: Matsui et al. proposed a text-prompted speaker recognition method, in which key sentences are completely changed every time the system is used [51]. The system accepts the input utterance only when it determines that the registered speaker uttered the prompted sentence. Because the vocabulary is unlimited, prospective impostors cannot know in advance the sentence they will be prompted to say. This method not only accurately recognizes speakers, but can also reject an utterance whose text differs from the prompted text, even if it is uttered by a registered speaker. Thus, a recorded and played back voice can be correctly rejected.

(3) ***Score normalization***: How to normalize intra-speaker variation of likelihood (similarity) values is one of the most difficult problems in speaker verification. Variations arise from the speaker him/herself, from differences in recording and transmission conditions, and from noise. Speakers cannot repeat an utterance precisely the same way from trial to trial. Likelihood ratio- and *a posteriori* probability-based techniques were investigated [28, 52, 66]. In order to reduce the computational cost for calculating the normalization term, methods using "cohort speakers" or a "world model" were proposed.

(4) ***Relation with other speech research***: Speaker characterization techniques are related to research on improving speech recognition accuracy by speaker adaptation [22], improving synthesized speech quality by adding the natural characteristics of voice individuality, and converting synthesized voice individuality from one speaker to another. Studies on automatically extracting the speech periods of each person separately from a dialogue/conversation/meeting involving more than two people have appeared as an extension of speaker recognition technology [27, 75, 88]. Increasingly, speaker segmentation and clustering techniques have been used to aid in the adaptation of speech recognizers and for supplying metadata for audio indexing and searching.

3.4. 2000s

(1) ***Score normalization***: A family of new normalization techniques has recently been proposed, in which the scores are normalized by subtracting the mean and then dividing by standard deviation, both terms having been estimated from the (pseudo) imposter score distribution. Different possibilities are available for computing the imposter score distribution: Znorm, Hnorm, Tnorm, Htnorm, Cnorm and Dnorm [6]. The state-of-the-art text-independent speaker verification techniques associate one or several parameterization level normalizations (CMS, feature variance normalization, feature warping, etc.) with a world model normalization and one or several score normalizations.

(2) ***High-level features***: High-level features such as word idiolect, pronunciation, phone usage, prosody, etc. have been successfully used in text-independent speaker verification. Typically, high-level-feature recognition systems produce a sequence of symbols from the acoustic signal and then perform recognition using the frequency and co-occurrence of symbols. In Doddington's idiolect work [13], word unigrams and bigrams from manually transcribed conversations were used to characterize a particular speaker in a traditional target/background likelihood ratio framework.

4. Discussions

4.1. Summary of the technology progress

In the last 50 years, research in speech and speaker recognition has been intensively carried out worldwide, spurred on by advances in signal processing, algorithms, architectures, and hardware. The technological progress in the 50 years can be summarized by the following changes [24]:

(1) from template matching to corpus-base statistical modeling, e.g. HMM and *n-grams*,
(2) from filter bank/spectral resonance to cepstral features (cepstrum + Δcepstrum + ΔΔcepstrum),
(3) from heuristic time-normalization to DTW/DP matching,
(4) from "distance"-based to likelihood-based methods,
(5) from maximum likelihood to discriminative approach, e.g. MCE/GPD and MMI,
(6) from isolated word to continuous speech recognition,
(7) from small vocabulary to large vocabulary recognition,
(8) from context-independent units to context-dependent units for recognition,
(9) from clean speech to noisy/telephone speech recognition,
(10) from single speaker to speaker-independent/adaptive recognition,
(11) from monologue to dialogue/conversation recognition,
(12) from read speech to spontaneous speech recognition,
(13) from recognition to understanding,
(14) from single-modality (audio signal only) to multimodal (audio/visual) speech recognition,
(15) from hardware recognizer to software recognizer, and
(16) from no commercial application to many practical commercial applications.

Most of these advances have taken place in both the fields of speech recognition and speaker recognition. The majority of technological changes have been directed toward the purpose of increasing robustness of recognition, including many other additional important techniques not noted above.

Recognition systems have been developed for a wide variety of applications, ranging from small vocabulary keyword recognition over dialed-up telephone lines, to medium size vocabulary voice interactive command and control systems for business automation, to large vocabulary speech transcription, spontaneous speech understanding, and limited-domain speech translation.

Although we have witnessed many new technological promises, we have also encountered a number of practical limitations that hinder a widespread deployment of applications and services.

4.2. Changes since 1977

Table 1 shows the research level of ASR techniques in 1977 [5]. Most of the techniques categorized into C: "a long way to go", printed in bold-face, still even now have not been able to overcome problems preventing realization of goals. Table 2 shows a list of ASR problems in 1977. Roughly speaking, 16 problems out of 28, printed by bold-face, have not yet been solved.

Table 1: State-of-the-art of ASR techniques in 1977 (A: useful now; B: shows promise; **C**: a long way to go) [5]

Processing Techniques	State-of-the-Art
1) Signal conditioning	A, except speech enhancement (C)
2) Digital signal transformation	A
3) Analog signal transformation and feature extraction	A, except feature extraction (C)
4) Digital parameter and feature extraction	B
5a) Resynthesis	A
5b) Orthographic synthesis	**C**
6) Speaker normalization **Speaker adaptation** **Situation adaptation**	**C**
7) Time normalization	B
8) Segmentation and labeling	B
9a) Language statistics	**C**
9b) Syntax	B
9c) Semantics	**C**
9d) Speaker and situation pragmatics	**C**
10) Lexical matching	**C**
11) Speech understanding	B-**C**
12) Speaker recognition	A for speaker verification ; **C** for all others
13) System organization and realization	A-**C**
14) Performance evaluation	**C**

4.3. How to decrease the gap between machine and human speech recognition

It has been shown that human speech recognition performs much better than the state-of-the-art ASR systems. In most recognition tasks, human subjects produce one to two orders of magnitude less errors than machines [46]. There is now increasing interest in finding ways to bridge this performance gap. It seems clear now that current problems in speech recognition can not be solved with only data-driven top-down approaches. Recent research in human speech processing has shown that human beings actually perform speech recognition by integrating multiple knowledge sources from bottom up [1].

What we know about human speech processing is still very limited, and we have yet to witness a complete and worthwhile unification of the science and technology of speech. In 1994, Moore [53] presented the following 20 themes which he believed important to the greater understanding of the nature of speech and mechanisms of speech pattern processing in general:

(1) How important is the communicative nature of speech?
(2) Is human-human speech communication relevant to human-machine communication by speech?
(3) Speech technology or speech science? (How can we integrate speech science and technology?)
(4) Whither a unified theory?
(5) Is speech special?
(6) Why is speech contrastive?
(7) Is there random variability in speech?
(8) How important is individuality?
(9) Is disfluency normal?
(10) How much effort does speech need?
(11) What is a good architecture (for speech processes)?
(12) What are suitable levels of representation?
(13) What are the units?
(14) What is the formalism?
(15) How important are the physiological mechanisms?
(16) Is time-frame based speech analysis sufficient?
(17) How important is adaptation?
(18) What are the mechanisms for learning?
(19) What is speech good for?
(20) How good is speech?

After more than 10 years, we still do not have clear answers to these 20 questions.

Table 2: ASR problems in 1977 [5] (Bold-face indicates problems that have still not been solved.)

1) Detect speech in noise; speech/nonspeech.
2) Extract relevant acoustic parameters (poles, zeros, formant (transitions), slopes, dimensional representation, zero-crossing distributions).
3) Dynamic programming (nonlinear time normalization).
4) Detect smaller units in continuous speech (word/phoneme boundaries; acoustic segments).
5) Establish anchor point; scan utterance from left to right; start from stressed vowel, etc.
6) Stressed/unstressed.
7) Phonological rules.
8) Missing or extra added ("uh") speech sound.
9) Limited vocabulary and restricted language structure necessary; possibility of adding new words.
10) Semantics of (limited) tasks.
11) Limits of acoustic information only
12) Recognition algorithm (shortest distance, (pairwise) discriminant, Bayes probabilities).
13) Hypothesize-and-test, backtrack, feed forward.
14) Effect of nasalization, cold, emotion, loudness, pitch, whispering, distortions due to talker's acoustical environment, distortions by communication systems (telephone, transmitter-receiver, intercom, public address, face masks), nonstandard environments.
15) Adaptive and interactive quick learning.
16) Mimicking; uncooperative speaker(s).
17) Necessity of visual feedback, error control, level for rejections.
18) Consistency of references.
19) Real-time processing.
20) Human engineering problem of incorporating speech understanding system into actual situations.
21) Cost-effectiveness.
22) Detect speech in presence of competing speech.
23) Economical ways to adding new speakers to system.
24) Use of prosodic information.
25) Coarticulation rules.
26) Morphology rules.
27) Syntax rules.
28) Vocal-tract modeling.

5. Conclusion

Speech is the primary, and the most convenient means of communication between people. Whether due to technological curiosity to build machines that mimic humans or desire to automate work with machines, research in speech and speaker recognition, as a first step toward natural human-machine communication, has attracted much enthusiasm over the past five decades. Although many important scientific advances have taken place, bringing us closer to the "Holy Grail" of automatic speech recognition and understanding by machine, we have also encountered a number of practical limitations which hinder a widespread deployment of application and services. In most speech recognition tasks, human subjects produce one to two orders of magnitude less errors than machines. There is now increasing interest in finding ways to bridge such a performance gap. What we know about human speech processing is very limited. Significant advances in speech and speaker recognition are not likely to come solely from research in statistical pattern recognition and signal processing. Although these areas of investigations are important, the significant advances will come from studies in acoustic-phonetics, speech perception, linguistics, and psychoacoustics. Future systems need to have an efficient way of representing, storing, and retrieving "knowledge" required for natural conversation [32]

6. References

[1] J. Allen, "From Load Rayleigh to Shannon: How do we decode speech?," *Proc. ICASSP*, 2002.

[2] B. S. Atal, "Text-independent speaker recognition," *J.A.S.A.*, 52, 181 (A), 83th ASA, 1972.

[3] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J.A.S.A.*, 55, 6, pp. 1304-1312, 1974.

[4] B. Beek, et al., "An assessment of the technology of automatic speech recognition for military applications," *IEEE Trans. Acoustics, Speech, Signal Processing*, ASSP-25, pp. 310-322, 1977.

[5] B. Beek, et. al., "Automatic speaker recognition system," *Rome Air Development Center Report*, 1971.

[6] F. J. Bimbot, et. al., "A tutorial on text-independent speaker verification," *EURASIP Journ. on Applied Signal Processing*, pp. 430-451, 2004.

[7] P. D. Bricker, et. al., "Statistical techniques for talker identification," *B.S.T.J.*, 50, pp. 1427-1454, 1971.

[8] J. S. Bridle and M. D. Brown, "Connected word recognition using whole word templates," *Proc. Inst. Acoust. Autumn Conf.*, pp. 25-28, 1979.

[9] W. Chou, "Mimimum classification error (MCE) approach in pattern recognition," W. Chou and B.-H. Juang (Eds.) *Pattern Recognition in Speech and Language Processing*, CRC Press, pp. 1-49, 2003.

[10] Y. L. Chow, et. al. "BYBLOS, the BBN continuous speech recognition system," *Proc. ICASSP*, pp. 89-92, 1987.

[11] K. H. Davis, et. al., "Automatic recognition of spoken digits," *J.A.S.A.*, 24 (6), pp. 637-642, 1952.

[12] G. R. Doddington, "A method of speaker verification," *J.A.S.A.*, 49, 139 (A), 1971.

[13] G. R. Doddington, "Speaker recognition based on idiolectal differences between speakers," *Proc. Eurospeech*, pp. 2521-2524, 2001.

[14] W. Endress, et. al., "Voice spectrograms as a function of age," Voice Disguise and Voice Imitation, *J.A.S.A.*, 49, 6(2), pp. 1842-1848, 1971.

[15] J. Ferguson, Ed., *Hidden Markov models for speech*, IDA, Princeton, NJ, 1980.

[16] J. W. Forgie and C. D. Forgie, "Results obtained from a vowel recognition computer program," *J.A.S.A.*, 31 (11), pp. 1480-1489. 1959.

[17] D. B. Fry, "Theoretical aspects of mechanical speech recognition"; and P. Denes, "The design and operation of the mechanical speech recognizer at University College London," *J. British Inst. Radio Engr.*, 19, 4, pp. 211-229, 1959.

[18] S. Furui, "An analysis of long-term variation of feature parameters of speech and its application to talker recognition," *Electronics and Communications in Japan*, 57-A, pp. 34-41, 1974.

[19] S. Furui, et. al., "Talker recognition by long time averaged speech spectrum," *Electronics and Communications in Japan*, 55-A, pp. 54-61, 1972.

[20] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoustics, Speech, Signal Processing*, ASSP-29, pp. 254-272, 1981.

[21] S. Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoustics, Speech, Signal Processing*, ASSP-34, pp. 52-59, 1986.

[22] S. Furui, "Speaker-independent and speaker-adaptive recognition techniques," in S. Furui and M. M. Sondhi, (Eds.) *Advances in Speech Signal Processing*, Marcel Dekker, pp. 597-622, 1991.

[23] S. Furui, "Speech-to-text and speech-to-speech summarization of spontaneous speech," *IEEE Trans. Speech & Audio Processing*, 12, 4, pp. 401-408, 2004

[24] S. Furui, "Fifty years of progress in speech and speaker recognition," *Proc. 148th ASA Meeting*, 2004

[25] S. Furui, "Recent progress in corpus-based spontaneous speech recognition," *IEICE Trans. Inf. & Syst.*, E88-D, 3, pp. 366-375, 2005

[26] M. J. F. Gales and S. J. Young, "Parallel model combination for speech recognition in noise," *Technical Report, CUED/F-INFENG/TR135*, 1993.

[27] H. Gish, M. Siu and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," *Proc. ICASSP*, S13.11, pp. 873-876, 1991.

[28] A. Higgins, et al., "Speaker verification using randomized phrase prompting," *Digital Signal Processing*, 1, pp. 89-106, 1991.

[29] F. Itakura, "Minimum prediction residual applied to speech recognition," *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-23 (1), pp. 67-72, 1975.

[30] F. Jelinek, et. al., "Design of a linguistic statistical decoder for the recognition of continuous speech," *IEEE Trans. Information Theory*, IT-21, pp. 250-256, 1975.

[31] F. Jelinek, "The development of an experimental discrete dictation recognizer," *Proc. IEEE*, 73 (11), pp. 1616-1624, 1985.

[32] B.-H. Juang and S. Furui, "Automatic speech recognition and understanding: A first step toward natural human-

machine communication," *Proc. IEEE*, 88, 8, pp. 1142-1165, 2000.

[33] B.-H. Juang and L. R. Rabiner, "Automatic speech recognition – A brief history of the technology development," K. Brown (Ed.) *Encyclopedia of Language and Linguistics*, Elsevier (to be published)

[34] B.-H. Juang and F. K. Soong, "Speaker recognition based on source coding approaches," *Proc. ICASSP*, 1, pp. 613-616, 1990.

[35] S. Katagiri, "Speech pattern recognition using neural networks," W. Chou and B.-H. Juang (Eds.) *Pattern Recognition in Speech and Language Processing*, CRC Press, pp. 115-147, 2003.

[36] T. Kawahara, et al., "Key-phrase detection and verification for flexible speech understanding," *IEEE Trans. Speech and Audio Proc.*, 6, 6, pp. 558-568, 1998.

[37] D. Klatt, "Review of the ARPA speech understanding project," *J.A.S.A.*, 62(6), pp. 1324-1366, 1977.

[38] C. H. Lee, "Statistical confidence measures and their applications," *Proc. ICSP*, pp. 1021-1028, 2001.

[39] C. H. Lee and L. R. Rabiner, "A frame synchronous network search algorithm for connected word recognition," *IEEE Trans. Acoustics, Speech, Signal Proc.,* 37 (11), pp. 1649-1658, November 1989.

[40] C. H. Lee, et. al., "Acoustic modeling for large vocabulary speech recognition," *Computer Speech and Language,* 4, pp. 127-165, 1990.

[41] K. F. Lee, et. al., "An overview of the SPHINX speech recognition system," *Proc. ICASSP,* 38, pp. 600-610, 1990.

[42] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, 9, pp. 171-185, 1995.

[43] K. P. Li and G. W. Hughes, "Talker differences as they appear in correlation matrices of continuous speech spectra," *J.A.S.A.*, 55, pp. 833-837, 1974.

[44] K. P. Li, et. al., "Experimental studies in speaker verification using a adaptive system," *J.A.S.A.*, 40, pp. 966-978, 1966.

[45] R. P. Lippmann, "An introduction to computing with neural nets," *IEEE ASSP Mag.*, 4 (2), pp. 4-22, April 1987.

[46] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, 22, pp. 1-15, 1997.

[47] Y. Liu, et. al., "Structural metadata research in the EARS program," *Proc. ICASSP*, V-957, 2005.

[48] B. Lowerre, "The HARPY speech understanding system," *Trends in Speech Recognition*, W. Lea, Ed., Speech Science Pub., pp. 576-586, 1990.

[49] T. B. Martin, et. al., "Speech recognition by feature abstraction techniques," *Tech. Report AL-TDR-64-176*, Air Force Avionics Lab, 1964.

[50] T. Matsui and S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs," *Proc. ICSLP*, pp. II-157-160, 1992.

[51] T. Matsui and S. Furui, "Concatenated phoneme models for text-variable speaker recognition," *Proc. ICASSP*, pp. II-391-394, 1993.

[52] T. Matsui and S. Furui, "Similarity normalization method for speaker verification based on a posteriori probability," *Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp. 59-62, 1994.

[53] R. K. Moore, "Twenty things we still don't know about speech," *Proc. CRIM/FORWISS Workshop on 'Progress and Prospects of Speech Research and technology',* 1994.

[54] C. S. Myers and L. R. Rabiner, "A level building dynamic time warping algorithm for connected word recognition," *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-29, pp. 284-297, 1981.

[55] K. Nagata, et. al., "Spoken digit recognizer for Japanese language," *NEC Res. Develop.*, 6, 1963.

[56] J. M. Naik, et. al., "Speaker verification over long distance telephone lines," *Proc. ICASSP*, pp. 524-527, 1989.

[57] H. F. Olson and H. Belar, "Phonetic typewriter," *J.A.S.A.,* 28 (6), pp. 1072-1081, 1956.

[58] D. B. Paul, "The Lincoln robust continuous speech recognizer," *Proc. ICASSP,* pp. 449-452, 1989.

[59] A. B. Poritz, "Linear predictive hidden Markov models and the speech signal," *Proc. ICASSP*, 2, pp. 1291-1294, 1982.

[60] S. Pruzansky, "Pattern-matching procedure for automatic talker recognition," *J.A.S.A.*, 35, pp. 354-358, 1963.

[61] S. Pruzansky and M. V. Mathews, "Talker recognition procedure based on analysis of variance," *J.A.S.A.*, 36, pp. 2041-2047, 1964.

[62] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE,* 77 (2), pp. 257-286, 1989.

[63] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliff, New Jersey, 1993.

[64] L. R. Rabiner, et. al., "Speaker independent recognition of isolated words using clustering techniques," *IEEE Trans. Acoustics, Speech, Signal Proc.,* ASSP-27, pp. 336-349, 1979.

[65] D. R. Reddy, "An approach to computer speech recognition by direct analysis of the speech wave," *Tech. Report No. C549*, Computer Science Dept., Stanford Univ., 1966.

[66] D. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp.27-30, 1994.

[67] R. Rose and R. A. Reynolds, "Text independent speaker identification using automatic acoustic segmentation," *Proc. ICASSP*, pp. 293-296, 1990.

[68] A. E. Rosenberg and M. R. Sambur, "New techniques for automatic speaker verification," *IEEE Trans. Acoustics, Speech, Signal Proc.,* ASSP-23, 2, pp. 169-176, 1975.

[69] A. E. Rosenberg and F. K. Soong, "Evaluation of a vector quantization talker recognition system in text independent and text dependent models," *Computer Speech and Language* 22, pp. 143-157. 1987.

[70] T. Sakai and S. Doshita, "The phonetic typewriter, information processing 1962," *Proc. IFIP Congress*, 1962.

[71] H. Sakoe, "Two level DP matching - a dynamic programming based pattern matching algorithm for connected word recognition," *IEEE Trans. Acoustics, Speech, Signal Proc.,* ASSP-27, pp. 588-595, 1979.

[72] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition,"

IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-26 (1), pp. 43-49, 1978.

[73] M. R. Sambur, "Speaker recognition and verification using linear prediction analysis," *Ph. D. Dissert.*, M.I.T., 1972.

[74] K. Shinoda and C. H. Lee, "A structural Bayes approach to speaker adaptation," IEEE Trans. Speech and Audio Proc., 9, 3, pp. 276-287, 2001.

[75] M. Siu, et. al., "An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers," *Proc. ICASSP*, pp. I-189-192, 1992.

[76] H. Soltau, et. al., "The IBM 2004 conversational telephone system for rich transcription," *Proc. ICASSP*, I-205-208, 2005.

[77] F. K. Soong, et. al., "A vector quantization approach to speaker recognition," *AT&T Technical Journal*, 66, pp. 14-26. 1987.

[78] M. Sugiyama, "Segment based text independent speaker recognition," *Proc. Acoust., Spring Meeting of Soc. Japan*, pp. 75-76 (in Japanese), 1988.

[79] J. Suzuki and K. Nakata, "Recognition of Japanese vowels - preliminary to the recognition of speech," *J. Radio Res. Lab,* 37 (8), pp. 193-212, 1961.

[80] C. C. Tappert, et. al., "Automatic recognition of continuous speech utilizing dynamic segmentation, dual classification, sequential decoding and error recovery," *Rome Air Dev. Cen, Rome, NY, Tech. Report TR-71-146*, 1971.

[81] N. Tishby, "On the application of mixture AR hidden Markov models to text independent speaker recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-30, 3, pp. 563-570, 1991.

[82] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," *Proc. ICASSP*, pp. 845-848, 1990.

[83] V. M. Velichko and N. G. Zagoruyko, "Automatic recognition of 200 words," *Int. J. Man-Machine Studies,* 2, pp. 223, 1970.

[84] T. K. Vintsyuk, "Speech discrimination by dynamic programming," *Kibernetika,* 4 (2), pp. 81-88, 1968.

[85] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimal decoding algorithm," *IEEE Trans. Information Theory*, IT-13, pp. 260-269, 1967.

[86] A. Weibel, et. al., "Phoneme recognition using time-delay neural networks," *IEEE Trans. Acoustics, Speech, Signal Proc.,* 37, pp. 393-404, 1989.

[87] M. Weintraub et al., "Linguistic constraints in hidden Markov model based speech recognition," *Proc. ICASSP,* pp. 699-702, 1989.

[88] L. Wilcox, et. al., "Segmentation of speech using speaker identification," *Proc. ICASSP*, pp. I-161-164, 1994.

[89] V. Zue, et. al., "The MIT summit speech recognition system, a progress report," *Proc. DARPA Speech and Natural Language Workshop,* pp. 179-189, 1989.

[90] G. Zweig, *Speech recognition with dynamic Bayesian networks*, Ph.D. Thesis, University of California, Berkeley, 1998.