An Introduction to Speech and Speaker Recognition

Richard D. Peacocke and Daryl H. Graf
Bell-Northern Research

Being able to speak to your personal computer, and have it recognize and understand what you say, would provide a comfortable and natural form of communication. It would reduce the amount of typing you have to do, leave your hands free, and allow you to move away from the terminal or screen. You would not even have to be in the line of sight of the terminal. It would also help in some cases if the computer could tell *who* was speaking.

If you want to use voice as a new medium on a computer workstation, it is natural to explore how speech recognition can contribute to such an environment. Here, we will review the state of speech and speaker recognition, focusing on current technology applied to personal workstations.

Limited forms of speech recognition are available on personal workstations. Currently there is much interest in speech recognition, and performance is improving. Speech recognition has already proven useful for certain applications, such as telephone voice-response systems for selecting services or information, digit recognition for cellular phones, and data entry while walking around a railway yard or clambering over a jet engine during an inspection.

Nonetheless, comfortable and natural communication in a general setting (no constraints on what you can say and how you say it) is beyond us for now, posing a problem too difficult to solve. Fortunately, we can simplify the problem to allow the creation of applications like the examples just mentioned. Some of these simplifying constraints are discussed in the next section.

Speech recognition, the ability to identify spoken words, and speaker recognition, the ability to identify who is saying them, are becoming commonplace applications of speech processing technology.

Speaker recognition is related to work on speech recognition. Instead of determining what was said, you determine who said it. Deciding whether or not a particular speaker produced the utterance is called *verification*, and choosing a person's identity from a set of known speakers is called *identification*. The most general form of speaker recognition (text-independent) is still not very accurate for large speaker populations, but if you constrain the words spoken by the user (text-dependent) and do not allow the speech quality to vary too wildly, then it too can be done on a workstation.

See the sidebar "Applications" for a description of typical speech and speaker recognition applications.

Factors affecting speech recognition

Modern speech recognition research began in the late 1950s with the advent of the digital computer. Combined with tools to capture and analyze speech, such as analog-to-digital converters and sound spectrograms, the computer allowed researchers to search for ways to extract features from speech that allow discrimination between different words. The 1960s saw advances in the automatic segmentation of speech into units of linguistic relevance (such as phonemes, syllables, and words) and on new pattern-matching and

0018-9162/90/0800-0026$01.00 © 1990 IEEE

Authorized licensed use limited to: UNIV OF MIAMI LIB. Downloaded on July 06,2023 at 09:48:43 UTC from IEEE Xplore. Restrictions apply.

Universal Electronics Inc. Exhibit 2005, Page 1 of 8
Roku, Inc. v. Universal Electronics Inc., IPR2022-01289

Applications

Although the performance of speech and speaker recognition systems is far from perfect, these systems have already proven their usefulness for certain applications.

Speech recognition. Currently, speech recognition is most often applied in manufacturing for companies needing voice entry of data or commands while the operator's hands are otherwise occupied. Related applications occur in product inspection, inventory control, command/control, and material handling. Speech recognition also finds frequent application in medicine, where voice input can significantly accelerate the writing of routine reports.

Speech recognition over the telephone network, although less used, has the greatest potential for growth. Automating the telephone operator's job can greatly reduce operating costs for telephone companies. Furthermore, speech recognition can help users control the personal workstation or interact with other applications remotely when touch-tone keypads are not available. (Telephone network applications are described in articles by Matthew Lennig and Ryohei Nakatsu elsewhere in this issue.)

Finally, speech recognition offers greater freedom to the physically handicapped.

Typical real-world applications:

- Delco electronics employs IBM PC/AT-Cherry Electronics and Intel RMX86 recognition systems to collect circuit board inspection data while the operator repairs and marks the boards.
- Southern Pacific Railway inspectors now routinely use a PC-based Votan recognition system to enter car inspection information from the field by walkie-talkie.
- Michigan Bell has installed a Northern Telecom recognition system to automate collect and third-number billed calls. AT&T has also put in field trial systems to automate call-type selection in its Reno, Nevada, and Hayward, California, offices.

Speaker recognition. Speaker recognition has been applied most often as a security device to control access to buildings or information. One of the best known examples is the Texas Instruments corporate computer center security system. Security Pacific has employed speaker verification as a security mechanism on telephone-initiated transfers of large sums of money. In addition to adding security, verification is advantageous because it reduces the turnaround time on these banking transactions. Bellcore uses speaker verification to limit remote access of training information to authorized field personnel. Speaker recognition also provides a mechanism to limit the remote access of a personal workstation to its owner or a set of registered users.

In addition to its use as a security device, speaker recognition could be used to trigger specialized services based on a user's identity. For example, you could configure an answering machine to deliver personalized messages to a small set of frequent callers.

classification algorithms. By the 1970s, a number of important techniques essential to today's state-of-the-art speech recognition systems had emerged, spurred on in part by the Defense Advanced Research Projects Agency speech recognition project. These techniques have now been refined to the point where very high recognition rates are possible, and commercial systems are available at reasonable prices.

Five factors can be used to control and simplify the speech recognition task[1]:

(1) *Isolated words.* Speech consisting of isolated words (short silences between the words) is much easier to recognize than continuous speech because word boundaries are difficult to find in continuous speech. Also, coarticulation effects in continuous speech cause the pronunciation of a word to change depending on its position relative to other words in a sentence. For example, "did you?" is not the same as "did" + short silence + "you?" Other effects depend on the rate of speaking as well, such as our tendency to drop the "t" in want when saying "want to" casually and quickly.

Error rates can definitely be reduced by requiring the user to pause between each word. For example, in a study by Bahl et al.,[2] error rates of 9 percent for continuous recognition decreased to 3 percent for isolated-word recognition. However, this type of restriction places a burden on the user and reduces the speed with which information can be input to the system (from a range of about 150-250 words per minute down to about 20-100 words per minute).

(2) *Single speaker.* Speech from a single speaker is also easier to recognize than speech from a variety of speakers because most parametric representations of speech are sensitive to the characteristics of the particular speaker. This makes a set of pattern-matching templates for one speaker perform poorly for another speaker. Therefore, many systems are speaker dependent — trained for use with each different operator. Relatively few speech recognition systems can be used by the general public. A rule of thumb used by many researchers is that, for the same task, speaker-dependent systems will have error rates roughly three to five times smaller than speaker-independent ones.

One way to make a system speaker independent is simply to mix training templates from a wide variety of speakers. A more sophisticated approach will attempt to look for phonetic features that are relatively invariant between speakers.

(3) *Vocabulary size.* The size of the vocabulary of words to be recognized also strongly influences recognition accuracy. Large vocabularies are more likely to contain ambiguous words than small vocabularies. Ambiguous words are those whose pattern-matching templates appear similar to the classification algorithm used by the recognizer. They are therefore harder to distinguish from each other. Of course, small vocabularies composed of many ambiguous words can be particularly difficult to recognize. A famous example is the E-set, which consists of a subset of the English alphabet and digits: "B," "C,"

Authorized licensed use limited to: UNIV OF MIAMI LIB. Downloaded on July 06,2023 at 09:48:43 UTC from IEEE Xplore. Restrictions apply.

Universal Electronics Inc. Exhibit 2005, Page 2 of 8
Roku, Inc. v. Universal Electronics Inc., IPR2022-01289

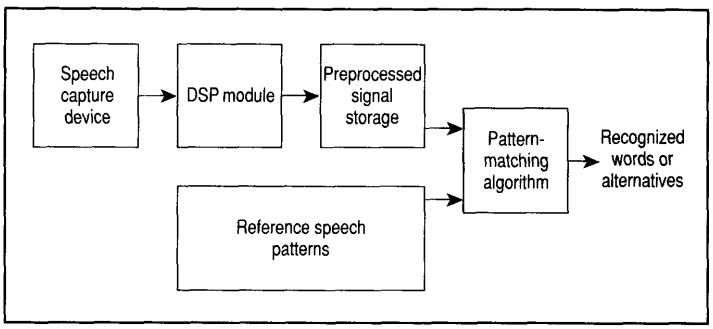


Figure 1. Components of a typical speech recognition system.

"D," "E," "G," "P," "T," "V," "Z," and "three."

The amount of time it takes to search the speech model database also relates to vocabulary size. Systems containing many pattern templates typically require pruning techniques to cut down the computational load of the pattern-matching algorithm. By ignoring potentially useful search paths, pruning heuristics can also introduce recognition errors.

(4) *Grammar*. The grammar of the recognition domain defines the allowable sequences of words. A tightly constrained grammar is one in which the number of words that can legally follow any given word is small. The amount of constraint on word choice is referred to as the *perplexity* of the grammar. Systems with low perplexity are potentially more accurate than those that give the user more freedom because the system can limit the effective vocabulary (and search space) to those words that can occur in the current input context. For example, a system described in Kimbal et al.[3] had an error rate of 1.6 percent with perplexity 19 (tightly constrained), while the error rate hit about 4.5 percent with perplexity 58 (more loosely constrained).

(5) *Environment*. Background noise, changes in microphone characteristics, and loudness can all dramatically affect recognition accuracy. Many recognition systems are capable of very low error rates as long as the environmental conditions remain quiet and controlled. However, performance degrades when noise is introduced or when conditions differ from the training session used to build the reference templates. To compensate, the user must almost always wear a head-mounted, noise-limiting microphone with the same response characteristics as the microphone used during training.

Components of a speech recognition system

Most computer systems for speech recognition include the following five components (see Figure 1):

(1) A *speech capture device*. This usually consists of a microphone and associated analog-to-digital converter, which digitally encodes the raw speech waveform.

(2) A *digital signal processing module*. The DSP module performs endpoint (word boundary) detection to separate speech from nonspeech, converts the raw waveform into a frequency domain representation, and performs further windowing, scaling, filtering, and data compression.[4] The goal is to enhance and retain only those components of the spectral representation that are useful for recognition purposes, thereby reducing the amount of information that the pattern-matching algorithm must contend with. A set of these speech parameters for one interval of time (usually 10-30 milliseconds) is called a *speech frame*.

(3) *Preprocessed signal storage*. Here, the preprocessed speech is buffered for the recognition algorithm.

(4) *Reference speech patterns*. Stored reference patterns can be matched against the user's speech sample once it has been preprocessed by the DSP module. This information is stored as a set of speech templates or as generative speech models.

(5) A *pattern matching algorithm*. The algorithm must compute a measure of goodness-of-fit between the preprocessed signal from the user's speech and all the stored templates or speech models. A selection process chooses the template or model (possibly more than one) with the best match.

Two major types of pattern matching in use are template matching by dynamic time warping and hidden Markov models. Artificial neural networks applied to speech recognition have also had some success, but this work is still in the early stages of research.[5] Moreover, linguistic knowledge incorporated into the pattern-recognition algorithm can enhance performance. However, such sophisticated techniques lie outside of the scope of this article (see, for example, O'Shaughnessy[4] and Mariani[6]).

Template matching by dynamic time warping became very popular in the 1970s. Template matching is conceptually simple. You want to compare the preprocessed speech waveform directly against a reference template by summing the distances between respective speech frames. However, biological limitations tend to produce nonlinear variations in timing from utterance to utterance. Consequently, the various frames of a word may be out of alignment with the corresponding frames of the given template. Since the order of speech events is fairly constant, you correct the misalignment by stretching the template in some places and compressing it in others to find an optimum match. Dynamic programming helps compute the optimum match. The sidebar "Dynamic time warping" illustrates the resulting time warp process.

Hidden Markov models are used in most current research systems because this technique produces better results for continuous speech with moderate-size vocabularies. HMMs are stochastic state machines that associate probabilities of producing sounds with transitions from state to state. An ideal HMM models speech with the same variations that occur in human speech due to coarticulation and other effects. Speech generated by a human being is matched against an HMM by computing the probability that the HMM would have generated the same utterance or by finding the state sequence through the HMM that has the highest probability of producing the utterance. The fact that HMMs generate poor-quality speech explains why recognition based on HMMs is still not perfect.

The sidebar "Hidden Markov models" further details the use of HMMs. Markov chains, although known about for almost a century, have only been successfully used in the context of speech recognition for the past 15 years or so. Until recently, no method existed for optimizing the model

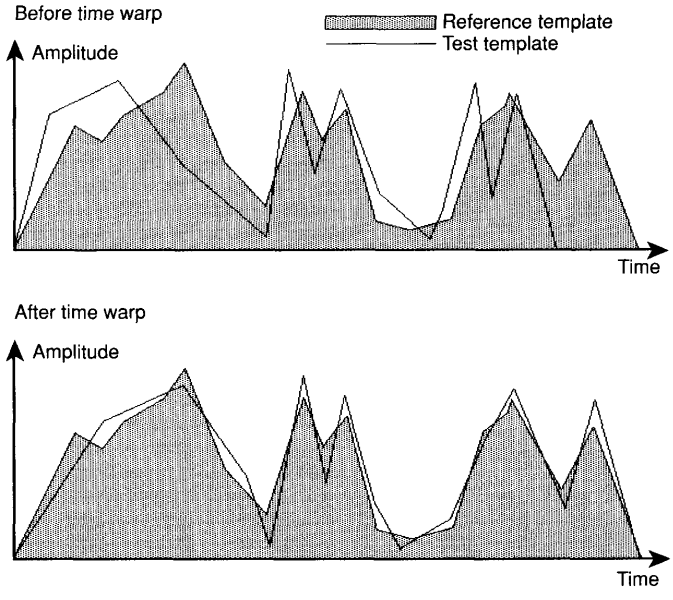
Authorized licensed use limited to: UNIV OF MIAMI LIB. Downloaded on July 06,2023 at 09:48:43 UTC from IEEE Xplore. Restrictions apply.

Universal Electronics Inc. Exhibit 2005, Page 3 of 8
Roku, Inc. v. Universal Electronics Inc., IPR2022-01289

Dynamic time warping

Frame distances between the processed speech frames and those of the reference templates are summed to provide an overall distance measure of similarity. But, instead of taking frames that correspond exactly in time, you would do a time "warp" on the utterance (and scale its length) so that similar frames in the utterance line up better against the reference frames. A dynamic programming procedure finds a warp that minimizes the sum of frame distances in the template comparison. The distance produced by this warp is chosen as the similarity measure.

In the illustration here, the speech frames that make up the test and reference templates are shown as scalar amplitude values plotted on a graph with time as the *x* axis. In practice, they are multidimensional vectors, and the distance between them is usually taken as the Euclidean distance. The graphs show how warping one of the templates improves the match between them. (For further information, see chapter 10 of O'Shaughnessy.[4])



parameters to generate observed speech patterns. (The US Department of Defense actually suppressed publication of the advances in HMM algorithms for a while in the mid-1970s, probably because of their use in cryptanalysis.) As well as representing low-level speech segments and transitions, hidden Markov models provide a framework on which you can model higher level structures in continuous speech signals and incorporate other knowledge about the communication.

Current speech recognition systems

Current speech recognition systems can be categorized according to the types of constraint they place on the speech. At one end of the spectrum fall speaker-independent, continuous, unconstrained-grammar, large-vocabulary systems. These systems are still very much in the research stage.

Several systems among those representing the state of the art were trained and tested on the same speech data — the DARPA resource management database — and are easily compared. The DARPA resource management task involves queries and commands to a database of warships. The associated database consists of a 997-word vocabulary and grammars with various complexities. Sphinx, a recognizer developed at Carnegie Mellon University, has a maximum word-recognition accuracy of 93.7 percent for a grammar of perplexity 60 and 70.6 percent for a grammar of perplexity 997.[1] BBN's Byblos[7] and a system developed at Lincoln Labs[8] have word accuracies of 88.7 percent and 87.4 percent, respectively, for the perplexity 60 grammar (BBN's system requires about two minutes of speech to adapt to a particular speaker before reaching this level of performance). Texas Instruments* and Stanford Research Institute[9] have reported systems with 44.3 percent and 40.4 percent accuracy on the perplexity 997 grammar. These systems have considerably lower *sentence* accuracies.

Representative of the state of the art in speaker-dependent, isolated-word, large-vocabulary recognizers are systems like IBM's Tangora recognizer, which is capable of 97 percent accuracy for a 20,000-word vocabulary[10] and NEC's 97.5 percent accurate, 1,800-word system.[11]

A variety of other systems trade off constraints on the input speech for higher recognition accuracies. Among these are the AT&T Bell Labs telephone-grade, speaker-independent, connected-digit recognizer (98.5 percent accurate when the number of digits is known[12]) and a speaker-dependent version of BBN's Byblos, which measured 94.8 percent accurate on the perplexity 60 DARPA resource management task.

At the highly constrained speech end of the spectrum fall speaker-dependent, single-word, small-vocabulary recognition systems. A variety of such systems developed can achieve accuracies above 99 percent.

Various commercial systems have appeared for Sun workstations and IBM-compatible PCs over the past few years. Table 1 summarizes the capabilities, costs, and manufacturers' claimed accuracies of a sample of these commercial products. Although several companies advertise speaker-independent, continuous, large-

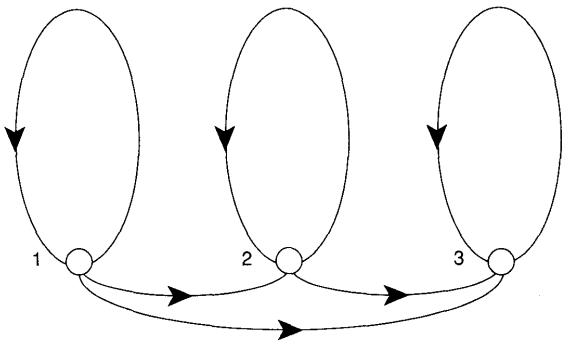
* See Kai-fu Lee,[1] p. 133.

Authorized licensed use limited to: UNIV OF MIAMI LIB. Downloaded on July 06,2023 at 09:48:43 UTC from IEEE Xplore. Restrictions apply.

Universal Electronics Inc. Exhibit 2005, Page 4 of 8
Roku, Inc. v. Universal Electronics Inc., IPR2022-01289

Hidden Markov models

A hidden Markov model (HMM) is a doubly stochastic process for producing a sequence of observed symbols. An underlying stochastic finite state machine (FSM) drives a set of stochastic processes, which produce the symbols. When a state is entered after a state transition in the FSM, a symbol from that state's set of symbols is selected probabilistically for output. The term "hidden" is appropriate because the actual state of the FSM cannot be observed directly, only through the symbols emitted. In the example here, the sequence of symbols AAaaB could have been produced by any of three different state transition sequences.



State	Possible Outputs
1	A,a
2	a
3	B

AAaaB could be produced by the following state sequences:

$\rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 3$
or $\rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 2 \rightarrow 3$
or $\rightarrow 1 \rightarrow 1 \rightarrow 2 \rightarrow 2 \rightarrow 3$

Although not shown in the example, probabilities are attached to the finite state transitions, and discrete probability distributions control the symbol output for each state (continuous density HMMs also exist). In the case of isolated word recognition, each word in the vocabulary has a corresponding HMM. These HMMs might actually consist of HMMs that model subword units such as phonemes connected to form a single word-model HMM. In the case of continuous word recognition, a single HMM corresponds to the domain grammar. This grammar model is constructed from word-model HMMs. The observable symbols correspond to (quantized) speech frame measurements.

An algorithm known as the forward/backward (or Baum-Welch) algorithm finds a set of state transition probabilities and symbol output distributions for each HMM. This gradient descent algorithm uses training data to iteratively refine an initial (possibly random) set of model parameters such that the HMM is more likely to generate patterns from the training set.

After this initial training stage, a word or sentence to be recognized is spoken, and speech measurements are made that reduce the utterance to a sequence of symbols. In the case of isolated word recognition, the forward algorithm computes the probability that each word model produced the observed sequence of symbols — the model with the highest probability represents the recognized word. In the case of continuous recognition, the Viterbi algorithm finds the state transition path, through the grammar model, with the maximum likelihood of generating the set of measurements. The sequence of word models on this path corresponds to the recognized sentence. (For further information see "Introduction to Hidden Markov Models" by L.R. Rabiner and B.H. Juang, published in *IEEE Trans. Acoustics, Speech, and Signal Processing*, Jan. 1986, pp. 4-16.)

vocabulary speech recognition, they carefully avoid making strong claims about the accuracy of their products. With commercial systems, you typically get what you pay for. Products available for less than $1,000 US are isolated-word, small-vocabulary recognizers. Speaker-dependent, isolated-word, large-vocabulary recognizers for automated dictation are available for a few thousand dollars. You'll see an order of magnitude leap in price when you move to large-vocabulary, speaker-independent, continuous-speech recognizers.

Speaker recognition — the voice, not just the words

Speaker recognition is related to speech recognition. When the task involves identifying the person talking rather than what is said, the speech signal must be processed to extract measures of speaker variability instead of being analyzed by segments corresponding to phonemes or pieces of text one after the other. For speaker recognition, only one classification is made, based on part or all of an input test utterance. Although various studies have shown that certain acoustical features work better than others in predicting speaker identity, few recognizers examine specific sounds because of difficulties in phone segmentation and identification.

Both automatic speaker verification and speaker identification use a stored database of reference patterns (templates) for N known speakers. Both involve similar analysis and decision techniques. Verification is simpler because it only requires comparing the test pattern against one reference pattern and it involves a binary decision: Is there a good enough match against the template of the claimed speaker? The error rate for speaker identification can be much greater because it requires choosing which of the N voices known to the system best matches the test voice or "no match" if the test voice differs sufficiently from all the reference templates.

Comparing test and reference utterances for speaker identity is much simpler for identical underlying texts, as in text-dependent speaker recognition. With cooperative speakers you can apply speaker recognition straightforwardly by using the same words to train the system and then test it. This usually happens in verification, but speaker identification often requires

Authorized licensed use limited to: UNIV OF MIAMI LIB. Downloaded on July 06,2023 at 09:48:43 UTC from IEEE Xplore. Restrictions apply.

Universal Electronics Inc. Exhibit 2005, Page 5 of 8
Roku, Inc. v. Universal Electronics Inc., IPR2022-01289

Table 1. A sample of commercially available speech recognition systems (for IBM PCs unless otherwise indicated).

Recognizer	Constraints (Speaker/Speech/Vocabulary)	Price (US $)	Percent Word Accuracy (per the manufacturer)
Dragon Voice-Scribe 400	Speaker dependent Isolated-word recognition 400 words	$995	>95
Dragon Dictate	Speaker adaptive Isolated-word recognition 30,000 words	$9,000	>90
ITT VRS 1280/PC	Speaker dependent Continous-speech recognition 2,000 words	$9,000	>98
Phonetic Engine* (Speech Systems Inc.)	Speaker independent Continous-speech recognition 10,000-40,000 words	$10,500-$47,100	95
Telerec (Voice Control Systems)**	Speaker independent Connected-word recognition 50 words	$3,000	>98
Verbex series 5000, 6000, 7000	Speaker dependent Continuous-speech recognition 80-10,000 words	$5,600-$9,600	>99.5
Voice Card (Votan)	Speaker dependent or independent Continuous-speech recognition 300 words	$3,500	>99 (speaker dependent), 95 (speaker independent)
Voice Comm Unit (Fujitsu)	Speaker dependent Connected-word recognition 4,000 words	Only in Japan	99.9
Voice Master Key (Covox)	Speaker dependent Isolated-word recognition 64 words	$150	95-96
Voice Navigator† (Articulate Systems)	Speaker dependent Isolated-word recognition 1,000 words	$1,300	95
Voice Pro (Voice Processing Corp.)	Speaker independent Continuous-speech recognition 13 words	$5,000	97-99
Voice Report (Kurzweil AI)	Speaker dependent Isolated-word recognition 20,000 words	$18,900	98

* Available for Sun workstations ** VCS technology is used in Dialogic products † Available for Macintosh (based on Dragon Systems technology)

text-independent methods. Higher error rates for text-independent methods means you will need much more speech data both for training and testing.

Automatic speaker recognition by computer has been an active research area since the early 1960s. A 1962 paper introduced the spectrogram as a means of personal identification, and this stimulated a good deal of further research. The term "voiceprint" also appeared in that paper. Unfortunately, the analogy with fingerprint read-

Authorized licensed use limited to: UNIV OF MIAMI LIB. Downloaded on July 06,2023 at 09:48:43 UTC from IEEE Xplore. Restrictions apply.

Universal Electronics Inc. Exhibit 2005, Page 6 of 8
Roku, Inc. v. Universal Electronics Inc., IPR2022-01289

ing is incorrect. As pointed out by Doddington,[13] the spectrogram is a function of the speech signal, not of the physical anatomy of the speaker. The speech signal depends far more on the speaker's actions, themselves a complex function of many factors, than on the shape of the speaker's vocal tract. The term "voiceprint" is misleading.

Speaker recognition systems

Speaker recognition by computer has only had limited success to date in applications using free text (text independent). Nonetheless, text-independent recognition of speakers has become an increasingly popular area of research, particularly for applications such as forensic, intelligence gathering, and passive surveillance of voice circuits. Free-text recognition usually lacks control over conditions that influence system performance, including variability in the speech signal and distortions and noise in the communications channel. The recognition task faces multiple problems: unconstrained input speech, uncooperative speakers, and uncontrolled environmental parameters. This has made it necessary to focus on features and characteristics of speech unique to the individual.

Performance of text-independent systems has lagged behind that of text-dependent systems, as you might expect. However, Markel and Davis[14] achieved excellent results with a linguistically unconstrained database of unrehearsed speech. Using voice pitch and linear predictive coding (LPC) reflection coefficients in their model, they reached 2 percent identification error and 4 percent verification error rates for 40-second segments of input speech. Results were not nearly as good with shorter input speech segments, even though the system avoided operational problems of microphone degradation, acoustic noise, and channel distortion. In text-independent recognition of nine male speakers over a radio channel at Bolt Beranek and Newman, the best performance was a 30 percent error rate for input speech segments of about two seconds.[15]

Text-independent recognition seems mainly slated for unobtrusive surveillance of individuals. As mentioned earlier, text-independent speaker identification poses a difficult problem. The accuracy of state-of-the-art text-independent identification is low, and it requires continuous use of computing power. Text-dependent speaker verification has the greatest potential for practical application at the moment. A number of organizations have research and development programs in speaker verification, and Texas Instruments and AT&T Bell Labs have both made major efforts in this research area.

AT&T Bell Labs has concentrated on speaker recognition over telephone lines, which faces difficult problems of microphone and channel distortion. Speaker recognition over telephone lines opens up an enormous set of possible uses, such as identification for various kinds of transaction processing in banking, shopping, and database access.

AT&T Bell Labs started its automatic speaker verification system in 1970. Researchers there chose measurements that are largely insensitive to the phase and spectral amplitude distortions likely over telephone lines. In an early five-month operational simulation, the system showed a user rejection rate and impostor acceptance rate of about 10 percent initially for new users, dropping to about 5 percent for experienced users and fully adapted templates. A more recent system used over telephone lines has achieved error rates (rejection of true speakers and acceptance of impostors) of approximately 2 percent.[16]

Texas Instruments has applied speaker verification to control access to its corporate computer center.[13] The gross rejection rate of the operational system measured 0.9 percent, with a casual impostor acceptance rate of 0.7 percent. The system has been operational 24 hours a day for more than a decade. The verification step uses a comparison of dynamic features, and time alignment is established using a simplified form of dynamic time warping. Verification utterances are constructed randomly using a four-word fixed phrase structure, for example, "Proud Ben served hard." As well as indicating what the user should say, the voice prompt helps stabilize pronunciation because the user tends to say it in the same way as prompted. Because a single utterance cannot provide the high level of verification performance desired, the system employs a sequential decision using multiple phrases when necessary. In operation, the actual number of phrases used averages 1.6.

Several companies offer speaker verification products, and voiceprints have emerged again. The claim of a voiceprint technique comes from Ecco Industries of Danvers, Mass. Ecco hopes to market a $300 voice-recognition security device for consumers this year. The intention is to create a digital picture of an individual's vocal tract so that bad colds or mimics will not trip up the device. US Sprint is testing a phone card that can be used to recognize the caller's voice. Card holders initiate calls in the usual way by dialing an 800 number, then dialing a personal identification number. A voice prompt asks callers to speak their password. The system is intended to reduce credit card fraud, including elimination of card-swapping. Bellcore has developed a similar credit card, which stores the template of a spoken word as 6,400 bits. According to Bellcore, the system recognizes the speaker 99 percent of the time, but the error rate can double to 2 percent if the speaker has a cold.

Speech recognition technology has migrated from mini- and mainframe computers to workstations and personal computers, and applications are already running on them. As embedded digital signal processors become more prevalent on workstations (such as the Next computer), we expect to see much wider use of speech and speaker recognition.

Current applications depend on the use of various simplifying constraints that make speech recognition feasible, as discussed above. The dependence on them means that, although useful practical applications of speech recognition exist, we have not yet achieved comfortable and natural communication with computers through voice. Text-dependent speaker recognition exists in the form of operational systems, but accurate text-independent speaker recognition remains a target.

Improvements in the speech and speaker recognition techniques discussed here will no doubt advance the performance of recognition systems, but it seems likely that we will also need natural language understanding before we can achieve comfortable and natural communication with computers through voice. ■

Acknowledgments

We are grateful for suggestions from the anonymous referees, which have improved the article, and to Jérôme Chiabaut and Taro Shibahara for help with an earlier version.

Authorized licensed use limited to: UNIV OF MIAMI LIB. Downloaded on July 06,2023 at 09:48:43 UTC from IEEE Xplore. Restrictions apply.

Universal Electronics Inc. Exhibit 2005, Page 7 of 8
Roku, Inc. v. Universal Electronics Inc., IPR2022-01289

References

1. K. Lee, *Automatic Speech Recognition: the development of the Sphinx System*, Kluwer Academic Publishers, Norwell, Mass., 1989.

2. L. R. Bahl et al., "Speech Recognition of a Natural Text Read as Isolated Words," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, April 1981, pp. 1,168-1,171.

3. D.O. Kimbal et al., "Recognition Performance and Grammatical Constraints," *Proc. DARPA Speech Recognition Workshop*, Feb. 1986, pp. 53-59.

4. D. O'Shaughnessy, *Speech Communication: Human and Machine*, Addison-Wesley, Reading, Mass., 1987.

5. M.A. Franzini, M.J. Witbrock, and K.-F. Lee, "Speaker-Independent Recognition of Connected Utterances Using Recurrent and Nonrecurrent Neural Networks," *Proc. Int'l Joint Conf. Neural Networks*, Vol.2, Washington, DC, June 1989, pp.II-1 to II-6.

6. J. Mariani, "Recent Advances in Speech Processing," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, Glasgow, Scotland, May 1989, pp. 429-440.

7. M.-W. Fung et al., "Improved Speaker Adaptation Using Text-Dependent Spectral Mappings," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, New York City, 1988, pp. 131-134.

8. D.B. Paul, "The Lincoln Robust Continuous Speech Recognizer," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, Glasgow, Scotland, 1989, pp. 449-452.

9. H. Murveit and M. Weintraub, "1,000-Word Speaker-Independent Continuous-Speech Recognition Using Hidden Markov Models," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, New York City, 1988, pp. 115-118.

10. W. Wylegala, "A 20,000-Word Recognizer Based on Statistical Evaluation Methods," *Speech Technology Magazine*, Apr./May 1989, pp. 16-18.

11. K. Yoshida, T. Watanabe, and S. Koga, "Large Vocabulary Word Recognition Based on a Demi-Syllable Hidden Markov Model Using a Small Amount of Training Data," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, Glasgow, Scotland, 1989, pp. 1-4.

12. L.R. Rabiner, J.G. Wilpon, and F.K. Soong, "High-Performance Connected-Digit Recognition, Using Hidden Markov Models," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, New York City, 1988, pp. 119-122.

13. G. R. Doddington, "Speaker Recognition — Identifying People by their Voices," *Proc. IEEE*, Vol. 73, No. 11, Nov. 1985, pp. 1,651-1,664.

14. J.D. Markel and B. Davis, "Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time-Spaced Database," *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol. ASSP-27, No. 1, 1979, pp. 74-82.

15. M. Krasner et al., "Investigation of Text-Independent Speaker Identification Techniques Under Conditions of Variable Data," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, 1984, Paper 18B.5.

16. M.R. Birnbaum, L.A. Cohen, and F.X. Welsh, "A Voice Password System for Access Security," *AT&T Tech. J.*, Vol. 65, No. 5, Sept./Oct. 1986, pp. 68-74.

Richard D. Peacocke is a member of the Computing Research Laboratory at Bell-Northern Research in Ottawa. At BNR he has worked on office communications and software engineering and has led groups working in software quality assurance and artificial intelligence. He has published a number of technical papers, including several on expert systems and telecommunications. His current research interests include AI, knowledge-based software, and speech technology.

Peacocke is president of the Canadian Society for Computational Studies of Intelligence and a member of the IEEE Computer Society, AAAI, and ACM. He is also an adjunct professor in the Department of Systems and Computer Engineering at Carleton University, Ottawa.

Peacocke has a PhD in computer science from the University of Toronto, Canada. He received a BA in mathematics from Cambridge University and an MS in computing science from the University of Alberta, Canada.

Daryl H. Graf is a member of the Computing Research Laboratory at Bell-Northern Research in Ottawa, Canada. Since joining BNR in 1979, he has acquired extensive experience in the telecommunications field. As a former manager of BNR's international operator services design group, he is interested in the application of speech processing technology to the operator services area. His current research interests include adaptive systems for pattern classification, and he has published technical papers on the use of neural networks for robot control.

Graf received the BS with honors in computer science from the University of Calgary in 1979 and the MCS with distinction from Carleton University, Ottawa, in 1988. He is a member of the IEEE, CSCSI, INNS, and AAAI.

Readers may contact the authors at the Computing Research Laboratory, Bell-Northern Research, PO Box 3511 Station C, Ottawa, Ontario, Canada K1Y 4H7, Bitnet: richard@bnr.ca or dgraf@bnr.ca.

Moving?

PLEASE NOTIFY US 4 WEEKS IN ADVANCE

Name (Please Print)

New Address

City State/Country Zip

MAIL TO:
IEEE Service Center
445 Hoes Lane
Piscataway, NJ 08854

ATTACH LABEL HERE

- This notice of address change will apply to all IEEE publications to which you subscribe.
- List new address above.
- If you have a question about your subscription, place label here and clip this form to your letter.

Authorized licensed use limited to: UNIV OF MIAMI LIB. Downloaded on July 06,2023 at 09:48:43 UTC from IEEE Xplore. Restrictions apply.

Universal Electronics Inc. Exhibit 2005, Page 8 of 8
Roku, Inc. v. Universal Electronics Inc., IPR2022-01289