Speech Recognition by Intelligent Machines

by *Hesham Tolba & Douglas O'Shaughnessy*
INRS-Télécommunications, Université du Québec, Montreal, QC

Abstract

The goal of Human-Machine (HM) technology is to create artificial intelligent machines that can interact with humans via voice. Explosive advances in the different fields of digital computing, signal processing and the evolution of statistical methods in the last ten years helped the huge progress and growth of HM research. However, the creation of such machines remains a distant goal. This is mainly due to the lack of a fundamental understanding of human speech processing. In this paper, we give an overview of human-machine technology with emphasis on Automatic Speech Recognition (ASR). Finally, we conclude with some perspectives about fundamental limitations in the current technology and some speculation about where we can go from here.

Sommaire

Le but de la technologie “interaction homme-machine” (HM) est de créer des machines artificielles intelligentes capables d'interagir avec les êtres humains par l'intermédiaire de leurs voix. Les progrès rapides dans les différents domaines tels que les calculs numériques, le traitement des signaux et l'évolution des méthodes statistiques au cours des dix dernières années ont contribué énormément au progrès et à la croissance énorme de la recherche sur la technologie HM. Cependant, la création de telles machines demeure encore un but éloigné. Ceci est principalement dû au manque d'une compréhension fondamentale du traitement de la parole par l'être humain. Dans cet article, nous donnons un aperçu globale de la technologie HM tout en mettant l'accent sur la reconnaissance automatique de la parole. Finalement, nous concluons avec quelques perspectives au sujet des limitations fondamentales de la technologie courante, et les axes de recherche les plus prometteurs pour améliorer cette technologie.

1.0 Introduction

The aim of human-machine technology is to create intelligent machines capable of decoding spoken information and acting appropriately upon that information, and then speaking to complete the information exchange [1]. However, the creation of such intelligent machines still remains a distant goal. Two challenging areas of speech research are still far from being mature enough to create such machines: automatic speech recognition (ASR) [2] and speech synthesis. For example, most existing ASR systems used for practical applications are of the small-vocabulary or isolated-word type. Medium- and large-vocabulary systems perform well in laboratories but not in real life [3] i.e. the field of ASR is in its early infancy. As well, synthesized speech is still far from natural speech [1]. The objective of this paper is to provide an overview of human-machine technology with emphasis on ASR.

2.0 ASR and its Difficulties

ASR can be described as the decoding of speech information using a machine (Figure 1). This decoded information can then be used to perform various tasks such as producing written text, controlling a machine or accessing a database, telephone voice dialing, and hands-free applications such as car phones. Huge progress in ASR research has occurred during the past four decades. However, the desired goal of a machine that can understand a spoken utterance on any subject by all speakers in different environments is still far from being achieved because of the associated difficulties. These difficulties include: inter- and intra-variability of speakers, the nature of the utterance (continuous speech versus isolated words), the vocabulary size, the complexity of the language and the robustness of such recognizers against different environmental conditions under which the recognition operation is performed. Although many of these problems have already been partially solved, there are still significant obstacles to be overcome before large-vocabulary continuous speech recognition systems can reach their full potential. In this section, we overview briefly the difficulties of ASR processes.

2.1 Adverse Conditions

A robust ASR system can deal with a broad range of applications and adapt to unknown conditions [6]. In general, the performance of existing speech recognition systems, whose designs are predicated on relatively noise-free conditions, degrades rapidly in the presence of adverse conditions. It was found that recognition accuracy for a typical speech recognizer drops from 96% for clean speech to 73% as the signal-to-noise ratio (SNR) is decreased to 20 dB, and it drops to 31% at 10 dB SNR. However, a recognizer can provide good performance even in very noisy background conditions if the exact testing condition is used to provide the training material from which the reference patterns of the vocabulary are obtained, which is practically not always the case.

In order to cope with the mismatched (adverse) conditions, different approaches could be used. Two fundamentally different approaches have been studied for achieving noise robustness. The first approach pre-processes the corrupted speech input signal prior to the pattern matching in an attempt to enhance the SNR. The second approach modifies the pattern matching itself in order to account for the effects of noise. Methods in this approach include noise masking, the use of robust distance measures, and HMM decomposition.

In addition to the above techniques, in certain applications, where training and testing can be done under the same noisy conditions, acceptable recognition performance can be obtained. It has been shown that this multi-style training improves the performance substantially under stress and with different speaking styles, under normal conditions by compensating for day-to-day speech variability. It can also be used when a recognizer cannot be trained under live stress conditions. Multistyle training reduces the error rate by more than a factor of two from 20.7% to 9.8%. The drop in the error rate is large, 6.2% to 2.9%, even for normally spoken words, and greatest for Lombard and angry conditions.

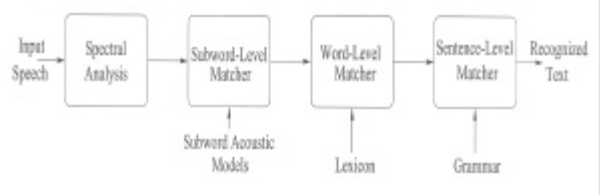

Figure 1: Block diagram of a typical continuos speech recognition system.

2.2 Inter- and Intra-speaker Variabilities

The inter- and intra-speaker variabilities in speech sounds include different speaking styles, speaking modes, diverse accents, poorly articulated speech, speaker stress, and disfluencies. Speaker noise includes the Lombard effect, uncooperative speakers, lip smacks, breath noises, pops, clicks, coughs, laughter, and sneezes. These variabilities result in two main categories of speech: read and spontaneous. These two types of speech differ not only in the way they are produced, but also in the way they are perceived. This was proven in different studies by showing that listeners can differentiate between the two speech types, even when lexical, syntactic, and semantic structure are identical. Although the perceptual distinction of the two types of speech is quite evident, it is not clear which perceptual cues enable such a distinction. However, in both cases, speakers include certain information in speech that enables listeners to recover words, and listeners apply what they know about the spoken language in order to understand such spoken speech. Several cues for word perception are used to recognize words. These cues include: the word itself, syntax and semantics, in addition to prosodic features.

2.3 Other Difficulties

In general, read speech is characterized by its monotony, fluency and correct syntax; however, spontaneous speech produces a rhythmic sensation and could be disfluent. Such disfluencies include: more hesitations and pauses, repetitions and repairs, false starts, pauses (either filled (vocalized), e.g. "uh", "um", etc., or lexicalized, e.g. "well", "like", and "you know", or unfilled (silent)), laughter and coughs, longer and nonuniform-distributed unfilled pauses. Beside disfluency, spontaneous speech is characterized by: pronunciation variation due to accents, coarticulation and speaking mode, phoneme deletion or phonemes shortened, less vowel reduction, sentence stress of some important words in terms of: pitch movement, variation of spectral characteristics (intensity), lengthening [10].

All of the above-mentioned disfluencies and the associated problems of spontaneous speech render the ASR process much more difficult and reduce the performance of recognizers.

3.0 Different Approaches for ASR

In general, there are two approaches to speech recognition, namely the acoustic-phonetic approach, and the pattern recognition-based approach. In the first approach, continuous speech can be segmented into well-defined regions which can then be given one of several phonetic labels based on measured properties of the speech features during the segmented region. Thus, characterization of the features of basic speech units can be found and speech can be labeled as a continuous stream of such phonetic units. Then, a mapping of the sequences of phonemic units into sequences of words is produced by the lexical decoding.

On the other hand, in the pattern recognition-based approach, the basic speech units are modeled acoustically based on a lexical description of words in the vocabulary. The acoustic-phonetic mapping is entirely learned via a finite training of a set of utterances. The resulting speech units are essentially acoustic descriptions of linguistically based units as represented in the words occurring in the training set. The pattern recognition-based phonemic approach has been found to have the highest recognition performance so far.

4.0 Main Components of an ASR System

Almost all speech recognition systems use a parametric representation to represent the waveform of a speech utterance. The aim of such a parameterization is: (1) to preserve the main features of speech that can easily identify a sound; (2) to eliminate as much as possible effects produced by communication channels, speaker differences and paralinguistic factors; and (3) to lower the information rate as much as possible for further easier processing, analysis and computation/memory reduction. A wide range of possibilities exists for parametrically representing the speech signal such as: the short-time spectral envelope, Linear Predictive coefficients (LPC), Mel-Frequency Cepstral Coefficients (MFCCs), the short-time energy, zero crossing rates and other related parameters (Figure 2).

To better represent temporal variations in the speech signal, higher-order time derivatives (or simply, delta parameters for first derivatives, delta-delta parameters for second derivatives) of signal measurements are added to the set of static parameters (e.g. MFCCs, LPC, etc.). The combination of dynamic and static features had proved additional discriminability for speech pattern comparison [2] and consequently improved the accuracy of the speech recognition process. Moreover, temporal variations in the speech signal, obtained by applying time derivatives to the speech signal, when combined with the static features mentioned above, had shown additional discriminability for speech pattern comparison [2].

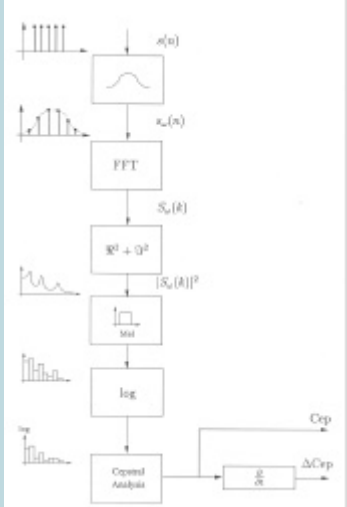

Figure 2: Front end speech parametrisation process.

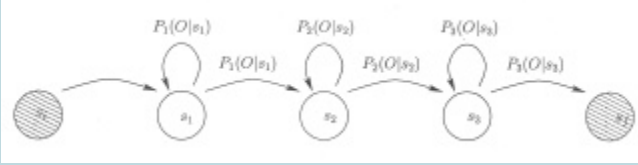

Figure 3: HMMs for context dependent phonemes

Once the feature vectors are formed, they are input to the pattern classification module. This module classifies these vectors into selected linguistic units (words, phonemes, etc.**). This can be achieved by Dynamic Time Warping (DTW), Hidden Markov Models (HMMs), Figure 3 [7], Artificial Neural Networks (ANNs) [8], expert systems [9] and combinations of such techniques. HMM-based systems are currently the most commonly used and most successful approach. This is followed by a word matcher, which takes the phonetic data from the front end and tries to make words out of them on the basis of stored phonetic rules, vocabulary, and syntax rules (i.e. linguistic constraints).

Sentences are formed by the concatenation of the words chosen by the word matcher according to syntax and semantic rules. To accomplish this task, a language model is used to provide a more accurate mechanism for estimating the probability of some word in an utterance given the preceding words. For simple tasks, in which it is only required to recognize a constrained set of phrases, we can use "rule-based" regular or context-free grammars. However, in large vocabulary tasks, "n-gram" (e.g. "bigram" and "trigram") grammars, with given probabilities of occurrence, are most commonly used. The word concatenation is done with an optional silence between words. These concatenated words are then matched to entire sentences which are stored in the lexicon, and the best matched sentence is selected. The sentence level matcher uses constraints imposed by a set of syntactic rules and semantic rules to determine the optimal sequence sentence in the language.

Searching for the best word sequence given the acoustic and language models and a spoken utterance is one of the most computational costs in a large vocabulary ASR system. Whatever the complexity of the acoustic models, the search cost is heavily influenced by the size of the vocabulary (task). Since the number of possible hypotheses grows exponentially with the length of the word sequence, the simple and obvious search strategy going through the whole HMM network is not practical in large-vocabulary ASR systems. In order to limit the exponential search space, heuristical pruning-away hypotheses with low scores techniques are used. The pruning causes the system to make suboptimal decisions while maintaining accuracy in the recognition process.

5.0 Statistical ASR Formulation

Statistical methods are the most dominant approach to speech recognition. Such popularity is due to their simplicity in modeling variations in speech signals using well-known statistical models such as Gaussian distributions, and training the systems using standard machine learning principles. Given an acoustic observation sequence:

$$O = \{o_1, o_2, \ldots, o_T\} \quad (1)$$

the ASR process is performed usually by finding the most likely sequence of acoustic phonetic units, $\hat{W}$, which maximizes the *a-posteriori* probability (MAP) as follows:

$$\hat{W} = \arg max_W \left[P(W|O)\right] \quad (2)$$

Since P(O) is independent of W then

$$\hat{W} = \arg max_W [P(O|W)P(W)] \quad (3)$$

*** Although the word is the natural unit to represent speech in a given language, large-vocabulary ASR systems require the modeling of speech in smaller units than words, i.e., subword units. This is due to the fact that the training of a large number of words is generally impractical. The subword unit used frequently in most existing ASR systems is the phone. There are, however, several variations of subword units. These variations include the context-dependent phone, e.g., biphones and triphones*

where P(O|W) is the probability of observing O when W is uttered.

If the phonetic unit used is the word, P(W) which is generally called the language model, it adds both syntactic and semantic constraints of the language to the recognition task. On the other hand, the MAP decoding can be accomplished by taking into account all possible sequences q of O:

$$\hat{W} = \underset{w,allq}{(\arg max)}\left[\sum_{qi \in q} P(W, q|O)\right] \quad (4)$$

6.0 State-of-the-art ASR Researches

The state-of-the-art ASR technology is the one that is based on HMM technology. That is, most research groups who work in the ASR field are concentrated in attacking the problems associated with HMM-based recognizers. These problems include the complexity of both acoustic and language models, the lexicon size, the searching algorithms and the robustness of such recognizers in adverse conditions. Some of the solutions to the above-mentioned problems that have been proposed over recent years include the use of huge amounts of data and the use of better smoothing techniques in order to refine the acoustic models. Complex language models are used instead of simple bigram and trigram ones. In addition, we use big lexicon and fast searching algorithms. Moreover, we concentrate also on rendering such recognizers more robust in adverse environments (against additive noise, convolutional noise (e.g. telephone speech), uncooperative speakers, rapid speech, spontaneous speech).

7.0 Problems with Existing Approaches to ASR

Almost all existing ASR systems use short-term parameter vectors representing about 10-20 ms of speech. The use of such short segments is inadequate to characterize phonemes in natural speech. In fact, the speech production phenomena of coarticulation, auditory phenomena of forward masking and the linguistic concept of a syllable point to temporal dynamics over an interval of several hundreds of milli-seconds. Thus, the use of short segments disregards all of these important acoustical aspects of speech.

As mentioned above, HMM-based systems are currently the most commonly used and most successful approach for ASR. HMMs became very popular models for ASR because they can deal efficiently with the temporal aspects of speech. In addition, there are powerful training and decoding algorithms that permit efficient training. Also, given their flexible topology, they can be extended to include some phonological or syntactic rules. To train these models, no explicit segmentation is needed, but only a lexical transcription, given a dictionary of phonological models, is necessary for the training of the HMMs.

However, HMMs do not exploit well many acoustical aspects of speech. For example, HMMs treat the very-short-term 10-ms frames of speech as separate information sources, i.e. correlation between successive acoustic vectors is not modeled well. This problem was solved partially by complementing the acoustic features that are used for ASR by their first and second time derivatives and/or using expensive linear discriminative training. Also, the assumption that the state sequences are first-order Markov chains, the prior choice of the model topology and the statistical distributions for each state disregard many acoustical aspects of speech. Besides, practical considerations such as numbers of parameters, the need of thousands of context-dependent phone models to handle coarticulation and the trainability of HMMs limit their implementations.

Another problem with existing ASR systems is the use of the MFCCs, which integrate the short-term spectral envelope of a speech signal over gradually wider intervals following the Mel scale. However, there is no theoretical basis that these coefficients are the optimal ones. In addition, it was found through experiments that these coefficients are highly sensitive to noise.

In addition to these problems, one fundamental problem for continuous speech recognition is the limitation of language models. As mentioned above, in large vocabulary tasks, “n-gram” (e.g. “bigram” and “trigram”) grammars, with given probabilities of occurrence, are most commonly used. n-grams can be estimated from simple frequency counts and stored in a look-up table. However, the problem is that the estimation of such trigrams is very poor due to the fact that many trigrams do not appear in the training data and many others will appear once or twice. To solve this problem, several models have been proposed, such as backing-off models, which are used when there are only one or two occurrences of trigrams in the training data. In such a case, backing-off is applied to replace the trigram probability by a scaled bigram probability. However, such models are very crude.

It must be noted also that all these models ignore hesitations, pauses, false starts, repetitions, etc. Thus, the problem remains unsolved especially for spontaneous speech.

Finally, one of the major drawbacks of existing ASR systems is their robustness against adverse conditions, as mentioned above. Although a lot of research has been conducted by most of the researchers who work on ASR to solve such a problem, it is still an open one, especially with the growing need for applications in wireless environments.

8.0 The Human Way versus the Machine Way

Studies have shown that the performance of existing recognizers are far short of the performance of humans in recognizing speech. This fact motivated several researchers to study the basic principles of human speech recognition (HSR) in an attempt to create artificially intelligent machines that are capable of mimicking humans in recognizing speech. In fact, both HSR and ASR have the same goal; i.e. to get the linguistic message from the signal. However, if we compare HSR manner to most existing ASR systems we find that human auditory perception works differently than current ASR systems. ASR machines use spectral matching techniques, but humans recognize speech with partial recognition of information across frequency [5]. That is, the linguistic message is independently decoded in different frequency subbands; the final decoding decision is based on merging the information from such bands. It was found that such an approach is effective as long as some sub-bands contain relatively uncorrupted information. That is, the information from the possibly corrupted sub-bands does not have to be used to decode the message. Thus, a better understanding of the partial recognition of speech processing in humans is required to get robust ASR. This approach was found effective when used for ASR if some sub-bands contained relatively uncorrupted information [4].

9.0 Future Research & Perspectives

Multilingual automatic speech recognition (ASR) in various acoustic environments is one of the most promising fields of speech communication research. Enormous progress in ASR research was made in the past 40 years. However, the desired goal of a machine which can understand a task-independent expression uttered by all speakers using various languages in different environments is still far from reality. Current research is now focused upon statistical methods. Improving the performance of ASR systems that are used to recognize spontaneous speech in adverse environments is still an open problem. This demand increases especially with the increase of the use of these systems in telephone applications.

Enhancing the performance of such recognizers in adverse environments can be achieved by using other auditory-based strategies instead of the Mel approximation in order to get more robust features that can be used for the recognition of both clean and telephone speech.

The recognition of spontaneous speech can be improved by taking into consideration the effects of the filled pauses while performing the recognition process by: (1) either omitting such pauses or by considering them as words to be added to the dictionary of the ASR system, (2) recognizing hesitations and restarts, (3) improving the model accuracy at both the acoustic level and at the language model level and (4) increasing the amount of training data and the lexicon size. This could reduce the error rate without increasing the search complexity.

Finally, we believe that a better understanding of the properties of human auditory perception that are relevant for decoding the speech signal and are likely to improve the performance of ASR in different environments is necessary for improving the performance of the existing recognizers. Also, using longer acoustic units (for example, syllables) instead of using short-term speech segments followed by post-processing techniques or using dynamic features is promising for the evolution of ASR. Moreover, rich prosodic cues (e.g., fundamental frequency (F0), energy, duration, etc.) that permit successful understanding, which are ignored by state-of-the-art ASR systems, must be considered for better performance. Also, the use of language-independent acoustic models and variable n-gram language models will enhance the performance further. Finally, we recommend strongly to benefit

from the results of the research of the HSR field by using a hybrid system that is not based only on statistical methods but also on speech communication knowledge. Using such a system could solve almost all the problems of ASR in the existing systems.

10.0 References

[1]. Douglas O'Shaughnessy, "Speech Communication: Human and Machine," IEEE Press, 2001.

[2]. L. Rabiner and B. H. Juang, "Fundamentals of Speech Recognition," Prentice Hall Inc., 1993.

[3]. J. Deller, J. Proakis and J. Hansen, "Discrete-Time Processing of Speech Signals," MacMillan Publishing Company, 1993.

[4]. H. Hermansky, "Should Recognizers have ears?", Speech Communication 25, pp. 3-27, 1998.

[5]. Jont B. Allen, "How Do Human Process and Recognize Speech?", IEEE Transactions on Speech and Audio Processing ASP, Vol. 2, No. 4, pp. 567-577, October, 1994.

[6]. Jean-Claude Junqua and Jean-Paul Haton, "Robustness in Automatic Speech Recognition," Kluwer Academic Publishers, 1996.

[7]. L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proc. of the IEEE, Vol. PROC-77, No. 2, pp. 257-286, February, 1989.

[8]. R. P. Lippmann, "Review of Neural Networks for Speech Recognition," Neural Computation, Vol. 1, No. 1, pp. 1-38, 1989.

[9]. Peter Jackson, "Introduction to Expert Systems," Addison-Wesley, 1999.

[10]. P. Howell and K. Kadi-Hanifi "Comparison of Prosodic Properties between Read and Spontaneous Speech", Speech Communication, 10, pp. 163-169, 1991.

About the authors

Hesham Tolba received a B.Sc. degree in electrical engineering and a M.Sc. degree in digital communications from Alexandria University in Egypt. He received his Ph.D. in Telecommunications from the INRS-Telecommunications, Quebec University, Québec, Canada. In 2000, he joined the INRS-Telecommunications as a professor in the speech communication group. He was previously a professor in the electrical engineering department at Alexandria University, Egypt. His current research interests include speech Processing, enhancement, synthesis and Recognition.

Douglas O'Shaughnessy has been a professor at INRS-Telecommunications (University of Quebec) in Montreal, Canada, since 1977. His interests include automatic speech synthesis, analysis, coding and recognition. He has been an Associate Editor for the Journal of the Acoustical Society of America since 1998, and will be the General Chair of the 2004 International Conference on Acoustics, Speech and Signal Processing (ICASSP) in Montreal, Canada.

The 2001 Canadian Engineers' Awards

May 26, 2001
New Brunswick Community College,
St. Andrews-by-the-Sea, N.B.

CONSEIL CANADIEN DES INGÉNIEURS
CANADIAN COUNCIL OF PROFESSIONAL ENGINEERS

Presented annually by CCPE since 1972, the Canadian Engineers' Awards recognize outstanding individual Canadian engineers, engineering projects, and teams of engineers. In 2001, CCPE will present its inaugural Award for the Support of Women in the Engineering Profession.

The theme for the 2001 Awards is Great Canadian Engineers: Connecting People, Connecting Worlds.

The 2001 Award recipients include an engineer whose goal is to make engineering an attractive career choice for Native Canadians, an inventor who is helping hard of hearing people to communicate, one of Chatelaine Magazine's Top 15 Canadian Women to Watch, an expert in highway safety who got his start as a teacher in a one-room school house, a university professor who is working hard to open the doors of engineering to women, one of Canada's leading researchers in oil sands technology, and a teacher who goes the extra mile to help students understand difficult concepts.

Photo features (from left to right) Noel Cleland, President of CCPE, and Dr. Charles Andrew Laszlo, C.M., O.B.C., P.Eng. winner of the Gold Medal Award for exceptional individual achievement and distinction in a field of engineering;

The 2001 winners of the Canadian Engineers' Awards are:

- **Don B. MacEwen**, P.Eng. - Medal for Distinction in Engineering Education for exemplary contribution to engineering teaching at a Canadian University,
- **Marc Lalande**, ing. - Meritorious Service Award for Community Service for exemplary voluntary contribution to a community organization or humanitarian endeavour,
- **Dr. Janet A.W. Elliott**, P. Eng. - Young Engineer Achievement Award for outstanding contribution in a field of engineering by an engineer 35 years of age or younger,
- **Dr. Frank R. Wilson**, P. Eng. - Meritorious Service Award for Professional Service for outstanding contribution to a professional, consulting or technical engineering association or society in Canada,
- **Dr. Nancy Mathis**, P. Eng. and **Marie Bernard**, ing. - Award for the Support of Women in the Engineering Profession for exemplary actions and contributions that open doors for women to successfully enter the engineering profession (Presented for the first time in 2001),
- **Dr. Charles Andrew Laszlo**, C.M., O.B.C., P.Eng. - Gold Medal Award for exceptional individual achievement and distinction in a field of engineering.