AUTOMATIC SPEECH RECOGNITION: A COMMUNICATION PERSPECTIVE

Bishnu S. Atal

AT&T Labs, Florham Park, NJ 07932

ABSTRACT

Speech recognition is usually regarded as a problem in the field of pattern recognition, where one first estimates the probability density function of each pattern to be recognized and then uses Bayes theorem to identify the pattern which provides the highest likelihood for the observed speech data. In this paper, we will take a different approach to this problem. In speech recognition, the goal is communication of information by voice and we will discuss the basics of speech recognition from a communication perspective. The speech signal at the acoustic level has a bit rate of 64 kb/s but the underlying sound patterns have an information rate of less than 100 b/s. What is the role of this high bit rate at the acoustic level? We will discuss the principles of decoding patterns that are submerged in an ocean of seemingly irrelevant information.

1. INTRODUCTION

Significant progress has been made during the past several years in the field of automatic speech recognition and the speech recognition technology has advanced to a level where it is being used in many applications, such as telephone call automation, automatic transaction processing, and dictation. But, most of this success has come from developing applications that work only in specific tasks or speaking environments. We are still far from reaching the goal of creating applications where voice communication provides an easy-to-use interface to computers in the same manner in which two people talk to each other. Will our current research directions lead us to the above goal? What can be done to move faster towards this goal? In this paper, we take a fresh look at the problem of automatic speech recognition, examine critically the fundamental underpinnings of the present technology, and seek to provide a new way of approaching the solution of this important problem.

2. SPEECH RECOGNITION AS A PATTERN RECOGNITION PROBLEM

Many different approaches to automatic speech recognition have been proposed; these include acoustic-phonetic theory, statistical pattern recognition, and neural networks. The most successful approach so far has proven to be the one based on statistical pattern recognition [1]. The pattern recognition approach is illustrated by a simplified block diagram shown in Fig. 1.

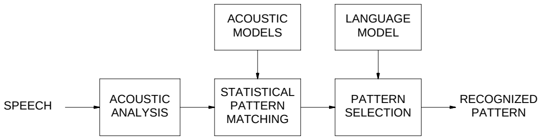

Fig. 1. Principal functions of a speech recognizer

In the first step, acoustic analysis is performed on the speech signal using a sequence of windows, resulting in a set of acoustic parameters once every few ms. Usually, some form of spectral analysis, providing a smooth envelope of the speech spectrum, is considered to be the preferred method of acoustic analysis. It is not clear that spectral analysis is the right choice to perform acoustic analysis. It is well known that intelligibility of speech produced by vocoders that use only spectral parameters is significantly lower than that of natural speech; only those speech coders that use information related to both excitation and spectral envelope are able to produce speech with high intelligibility. We will return to this point later in the paper.

In the second step, the set of acoustic parameters for the unknown speech are compared to a stored set of acoustic patterns derived from a large collection of labeled speech utterances from many speakers using an HMM-based training procedure. This comparison provides a set of likelihood scores representing the similarity between the unknown pattern and each of the stored patterns (or some combination there of).

Finally, in the last step, the likelihood scores are augmented with higher level knowledge about the speech utterance derived from a language model, the context, or task semantics, to produce the recognized pattern (or set of patterns) with the highest likelihood score.

Let us examine the second step, namely computation of likelihood score, in more detail. It is assumed that a specified word sequence **w** produces an acoustic parameter sequence **y** with a probability density function $P(\mathbf{w}|\mathbf{y})$. The problem of associating an arbitrary pattern **y** to a word sequence is solved by using Bayes' rule of conditional probability. The probability of error in recognizing a pattern is minimized if the recognized word string **w** is selected so that $P(\mathbf{w}|\mathbf{y})$ is maximum [2]. Using Bayes theorem, one can write

$$P(\mathbf{w}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{w})P(\mathbf{w})}{P(\mathbf{y})}.$$

The conditional density function $P(\mathbf{y}|\mathbf{w})$ is often called the *likelihood function*. In the pattern recognition approach, the decision problem is posed in probabilistic terms, assuming implicitly a complete knowledge of all relevant probabilities. As a practical matter, the conditional probability densities are not known. Usually, the unknown probability densities are estimated from a training set. In practice, this does not work well. Often, one tries to salvage this impossible situation by assuming that the form of the probability density is known (unrealistic assumption) and only some of the underlying parameters of the densities are unknown and can be estimated from the training data.

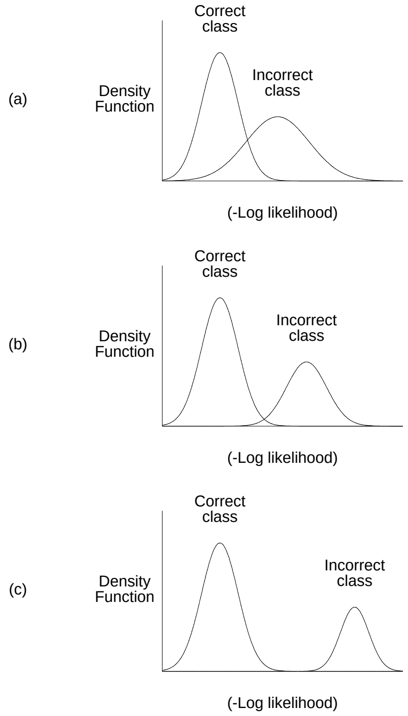

Fig. 2. Probability densities of log likelihood: (a) with significant overlap, (b) with small overlap, and (c) with almost no overlap between the densities for the correct and incorrect classes

The likelihood $P(\mathbf{y}|\mathbf{w})$ for a given $\mathbf{w}$ varies as a function of $\mathbf{y}$. Let us consider the probability density function of likelihood (or log likelihood). We will consider 3 cases. The first case is illustrated in Fig. 2 (a) which shows the density functions of the negative of log likelihood, namely $-\log P(\mathbf{y}|\mathbf{w})$, for the correct and the incorrect classes. In this case, there is a significant overlap between the two densities which will result in significantly inaccurate recognition of word sequences. That is not a desirable situation. Let us consider another example illustrated in Fig. 2 (b) where the overlap between the probability densities is much smaller. The recognition errors are contributed by the region where the density functions overlap; the non-overlapping parts of the probability density functions do not influence the errors. Since the overlapping region in Fig. 2 (b) is small, the shape of the density functions plays a minor role in determining the recognition errors. In the third case illustrated in Fig. 2 (c), there is almost no overlap between the two density functions and therefore the shape of the density functions is irrelevant in deciding about the identity of the unknown pattern. It is therefore interesting to note that the shapes of both the density functions of the log likelihood and $P(\mathbf{y}|\mathbf{w})$, which will produce little or no recognition errors are irrelevant. What matters is that the density functions of the log likelihood for the correct and incorrect classes are well separated and do not overlap.

For reliable speech recognition, the density functions of the likelihood must be narrow, although the acoustic patterns **y** for any **w** might be spread over a large region in the acoustic space. Therefore the central problem in speech recognition is not how to estimate $P(\mathbf{y}|\mathbf{w})$, but to ensure that the density functions of the likelihood are localized and do not overlap. The theory of statistical pattern recognition is built around the conditional probability density function of **y**, namely $P(\mathbf{y}|\mathbf{w})$, but what really matters is the probability density function of the likelihood P. The emphasis on estimating density functions P is completely misguided. The density functions $P(\mathbf{y}|\mathbf{w})$ are beyond our control, but the density functions of the log likelihood can be designed for accurate speech recognition. So far we discussed the statistical approach to speech recognition, but let us now look at the speech recognition problem from a communication perspective.

3. ASR: A COMMUNICATION PERSPECTIVE

Automatic speech recognition (ASR) is concerned with communication of information by voice: to communicate to a machine what a user wants the machine to do. Figure 3 shows ASR as part of a communication system. The signal x(t) represents the message to be transmitted. The message could be a string of text or in some other form. The significant point is that the actual message is the one selected from a set of possible messages. The message x(t) is transformed by the human speech production system to the speech signal s(t) which is transmitted on a communication channel to the speech recognizer where the signal s(t) is transformed to another signal y(t), the received message. In general, y(t) will not be identical to x(t), but the difference between the two must be kept small for proper communication.

Now here are the problems. First, the transformation F from x(t) to s(t) is not one fixed transformation but differs widely from one speaker to another. Second, the communication channel adds noise, reverberation, etc. and can introduce spectral distortions. The speech recognizer must be able to handle all this extra variability in transforming the speech signal to the received message y(t).

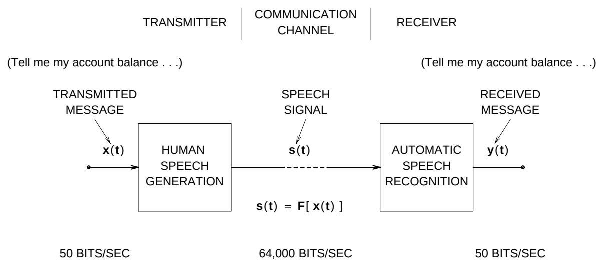

Fig. 3. Speech recognition: a communication perspective

The information rate of the speech signal s(t) is much larger than that of the message signal x(t) -- 64 kb/s for s(t) compared to about 50 b/s for x(t). This large amount of redundancy in the speech signal is necessary to cope with large speaker-to-speaker variability and the distortions introduced by the communication channel. Robustness in communication is always achieved by large expansion in the bandwidth, such as in FM, PCM, and CDMA systems. The speech signal is another example, where an extraordinary amount of robustness is accomplished by converting a message signal x(t) with a low information rate to a speech signal with a high information rate. The human vocal tract generates a speech signal with a bandwidth of 8 kHz, whereas the underlying sound patterns have a much lower bandwidth in the vicinity of 30 Hz or so. The task of the ASR is to extract a very small amount of information (about 50 b/s) that is relevant for identifying sound patterns in speech in the presence of a large amount of irrelevant information in the acoustic waveform. How can we determine the amount of information related to sound patterns that can be extracted from the acoustic signal?

4. Channel Capacity of Acoustic Parameters

We will obtain a crude estimate of the information that can be used to distinguish sound patterns in the speech signal. We will follow an approach similar to one used by Shannon in deriving the capacity of a channel in the presence of noise [3,4]. Let N represent the variance of ‘‘noise’’. The noise can be determined if we have a database of speech utterances, where each sound pattern is identified and labeled. We will first assume that the acoustic parameters are uncorrelated and the variance of each parameter is the same.

A speech segment of duration T and bandwidth B can be represented by $2BT$ parameters or by a point in $2BT$-dimensional hyperspace. Let $n = 2BT$. For a hypersphere of high dimensionality, almost all of the volume is very close to the surface. Therefore, for large values of n, the points for speech segments associated with the same sound pattern will lie very close to the surface of a hypersphere of radius $\sqrt{nN}$ and a speech vector corresponding to a sound pattern will be contained in a hypersphere of radius $\sqrt{nN}$ around a point representing that sound. Similarly, speech vectors corresponding to different sounds will be contained in a hypersphere of radius $\sqrt{n(S+N)}$, where S represents the variance of the ‘‘signal’’. The number of sound patterns that can be distinguished at the receiver equals the maximum number of non-intersecting hyperspheres of radius $\sqrt{nN}$ that can be placed in a hypersphere of radius $\sqrt{n(S+N)}$ in $2BT$ dimensions. The number is clearly $[(S+N)/N]^{n/2}$. The number of bits of information (channel capacity) is given by

$$C = (n/2)\log_2[(S+N)/N].$$

These sphere packing ideas are illustrated in a geometrical manner in Fig. 4.

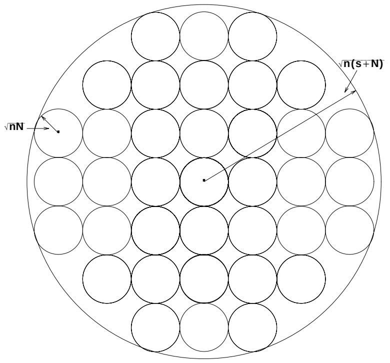

Fig. 4. Geometrical interpretation of sphere packing

Let us consider the case when both signal and noise variances are different for each acoustic parameter. Let S_k and N_k represent the signal and noise variances, respectively, for the kth parameter. Then the channel capacity is given by

$$C = (n/2)\sum_k \log_2[S_k + N_k]/N_k].$$

As an illustration, we apply now these ideas to determine the information conveyed by spectral parameters in distinguishing speech sounds. We have used a subset of DARPA TIMIT continuous speech database [5] which is available with all the phonetic segments labeled by human listeners. We have used 25 mel spectrum parameters as the acoustic parameters obtained by processing speech filtered to a bandwidth of 4 kHz using 20-ms Hamming window spaced at 10 ms intervals.

The full TIMIT database contains a total of 6300 utterances, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. We selected a subset of the database corresponding to the northern dialect region, which consists of 1020 utterances spoken by 102 speakers (71 male, 31 female). We grouped the various phones into 40 classes. Each frame of mel spectral parameters represents a speech segment of 20 ms duration. We combined successive frames spaced at 10 ms intervals to produce a super-frame of parameters for speech segments with duration ranging from 20 ms (2 frames) to 130 ms (12 frames).

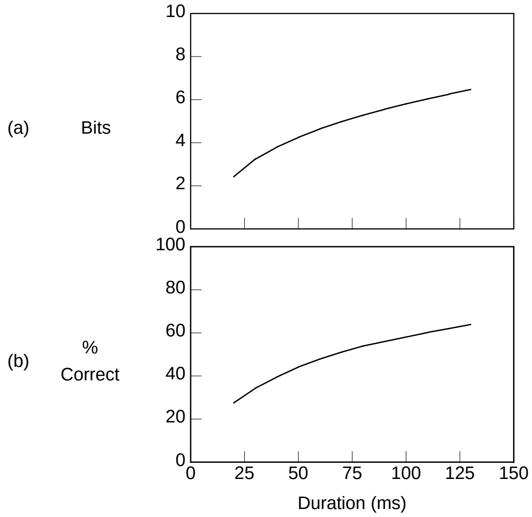

Fig. 5 (a). Information rate (bits) versus duration of segment
Fig. 5 (b). Recognition accuracy (%) versus duration of segment

The signal-to-noise ratio and the total number of bits of information were computed for each super-frame using the procedure outlined earlier for computing the channel capacity . These results are shown in Fig. 5 (a) which shows a plot of bits versus the duration of the speech segment in ms. The number of bits was 2.4 for a single spectral frame of 20 ms and 6.5 for a segment with a duration of 130 ms (12 spectral frames). We also determined the recognition accuracy as a function of duration. These results are shown in Fig. 5 (b). The percent recognition accuracy of a single spectral frame is 28 % for a segment of duration 20 ms and 64 % for a segment with a duration of 130 ms. Although it might appear that 6.5 bits of information (represented in the 130 ms long speech segment) should be sufficient for reliable recognition of 40 sound classes used in the study, a recognition accuracy of only 64 % was achieved. The reason for this result is that the channel capacity formulation is valid only as a limit when the number of dimensions $n = 2BT$ is infinitely large. In our case, the mel spectrum provided only 25 highly correlated parameters. The effective dimensionality in this case is far less than 25. The same holds for successive spectral frames that are spaced 10 ms apart.

6. DISCUSSION

In this paper, I have raised many issues concerning the application of statistical pattern recognition techniques to the problem of automatic speech recognition. These techniques are useful only when the various patterns cannot be distinguished, requiring the selection of the best choice according to some suitably chosen cost function. Speech recognition systems must be designed to be accurate, and recognition errors, if any, must be very few and sparse. The probability density functions of log likelihood for different classes must be non-overlapping to achieve accurate speech recognition. In such a situation, pattern recognition techniques that require careful estimation of probability densities for the observed acoustic parameters are irrelevant.

We have outlined an approach that makes it possible to determine how much information is available in the speech signal to distinguish between different sound patterns. Our results indicate that 25 spectral envelope parameters do not contain sufficient information to distinguish between various sounds in speech accurately, even when the parameters are combined over a time interval of 130 ms. To achieve accurate phone recognition, it will be necessary to move beyond spectrum envelope. As I indicated earlier, speech coders that used only the spectral information suffered from serious loss of intelligibility and it was found necessary to add fairly detailed information about the excitation in the form of multi-pulse excitation to produce speech that was close to the natural speech. I believe that in order to achieve accurate recognition, one has to use acoustic parameters that are closely related to the speech waveform rather than the spectrum. The goal should be to find a representation of speech that can provide 10 bits of information to distinguish between 40 speech sounds.

REFERENCES

[1] Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of speech recognition,* (Prentice Hall, Englewood Cliffs)

[2] Bahl, L. R., Jelenik, F., and Mercer, R. L. (1983). ‘‘A maximum likelihood approach to continuous speech recognition,’’ *IEEE Trans. Patt. Anal. Machine Intell.,* vol. PAMI-5, pp. 179-190.

[3] Shannon, C. E. (1949). ‘‘Communication in the presence of noise’’, *Proc. IRE,* vol. 37, pp. 10-21.

[4] Pierce, J. R. (1961). F2symbols, signals, and noise, Harper & Brothers, New York).

[5] Garofolo, J. S. (1988). ‘‘Getting started with the DARPA TIMIT CDROM: An acoustic phonetic continuous speech database,’’ National Institute of Standards and Technology (NIST), Gaithersburg, MD.