Pattern Recognition 41 (2008) 2965–2979

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Invited paper: Automatic speech recognition: History, methods and challenges☆

Douglas O'Shaughnessy

INRS-EMT, University of Quebec, 800 de la Gauchetiere West, Montreal, Quebec, Canada H5A 1K6

ARTICLE INFO

Article history:
Received 22 February 2008
Received in revised form 2 May 2008
Accepted 7 May 2008

Keywords:
Automatic speech recognition
Hidden Markov models
Adaptation
Compensation
Pattern recognition
Spectral representation

ABSTRACT

The field of automatic speech recognition (ASR) is discussed from the viewpoint of pattern recognition (PR). This tutorial examines the problem area, its methods, successes and failures, focusing on the nature of the speech signal and techniques to accomplish useful data reduction. Comparison is made with other areas of PR. Suggestions are given for areas of future progress.

1. Introduction

Among the tasks for which machines may simulate human behavior, automatic speech recognition (ASR) has been foremost since the advent of computers. The logical partner of ASR, automatic speech synthesis, existed before practical computing machines, although the quality of synthetic speech has only recently become reasonable. In earlier times, devices were built that approximated the acoustics of human vocal tracts (VTs), as the basic mechanisms of speech production were evident to early scientists, using models based upon musical instruments. A device to understand speech, however, needed a calculating machine capable of making complex decisions, and, practically, one that could function as rapidly as humans. As a result, ASR has grown roughly in proportion to other areas of pattern recognition (PR), in large part based on the power of computers to capture a relevant signal and transform it into pertinent information, i.e., recognizing a pattern in the (speech) signal.

As in any PR task, ASR seeks to understand patterns or ``information'' in an input (speech) waveform. For such tasks, an algorithm designer must estimate the nature of what ``patterns'' are sought. The target patterns in image PR, for example, vary widely: people, objects, lighting, etc. When processing audio signals such as speech, target information is perhaps less varied than video, but there is nonetheless a wide range of interesting patterns to distill from speech signals. The most common objective of ASR is a textual translation of the speech signal, i.e., the text corresponding to what one has said. Other useful outputs include: the language of the speech, the speaker's emotional state, and the speaker's identity [1]. A very practical use for ASR is as part (along with natural language understanding and automatic speech synthesis) of a human–machine dialogue, whereby a user can interact efficiently with a database, e.g., telephony [2].

Image and speech PR have both similarities and differences. In a sense, video has much greater variability than audio. Many images that meet the eye (or camera), whether natural or artificial (e.g., art, construction), vary greatly in their production, whereas the vast majority of sounds that meets the ear (or microphone) falls into a smaller set of categories. The latter include speech, music, animal sounds, machine sounds, and environmental sounds. In each of these audio classes, there are many features that help humans identify their sound source rapidly: periodicity, directionality, dynamic nature, spectral balance, etc. Such features can, of course, be exploited by machine PR, and we shall describe how this is done for ASR.

For speech, whether produced naturally by a human or reproduced by a machine, the sound origin (as typically assumed by a listener) is a speaker's VT. Thus, ASR has an input signal that is quite different from images, where input may be any display in the form of a gray-scale (or colored) pattern in two spatial dimensions (or in three dimensions for video, adding time as a variable). Human viewers of an image (or image sequence in time) usually try to impose or assume some physical ``structure,'' in terms of reference patterns, while trying to interpret the image, but the potential range of

☆ This work was supported by the Canadian government under the NSERC Strategic Partnerships Program.

E-mail address: dougo@emt.inrs.ca.

0031-3203/$30.00 © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.05.008

possibilities for images is indeed vast. For audio input, on the other hand, a listener will normally and readily label different parts of what they hear as coming from various elements of a limited set of classes (i.e., speech, music, etc). For speech specifically, the restrictions on possible sounds are significant; listeners will normally reject (as non-speech) any audio signal that could not have originated in a VT, in their experience of speech communication. When listening to speech, they assume a VT source and decipher the audio content in terms of what the speaker likely had in mind.

1.1. Variability in speech

While emphasizing the major difference in diversity between speech and various other signals (e.g., images) that are processed by humans, one must note nonetheless a large range of variability in speech signals [3]. Each person has a different VT, controlled by a unique brain. While speakers of any given language follow the same general linguistic rules, there is great latitude in how this is done, producing a vast range of ``acceptable'' utterances that would normally be properly interpreted by most listeners. It is impossible for humans to reproduce the same exact action twice; even when attempting to repeat a word uniformly, slight variations occur. These changes are readily observed in digital representations of speech signals.

Some ASR systems focus on a very limited number of speakers, e.g., subscribers to a service or purchasers of a specific ASR product. In such ``speaker-dependent'' (SD) cases, speech variations are typically less vast (vs. ``speaker-independent'' (SI) cases, where an ASR system makes no assumption of who is talking). However, even when speech is limited to one cooperative speaker, significant variations are often evident owing to environmental (e.g., different communication channels) and speaking conditions (e.g., words in different contexts). When we generalize the ASR task to be SI, as in most services for the general public, we face the much larger range of variability that arises from different people, with their varied VTs and diverse styles of speaking.

The biggest challenge for ASR is how to handle all this variability. As in any PR, a designer develops models or templates for signals of interest, from observed ``training'' data in an initial development phase, and then verifies the performance of the algorithm on new ``testing'' data. (As in all PR, it is essential to test on data not employed during training, as otherwise the risk is great that models would be ``over-trained'' toward the data they have already seen, and thus under-generalized for future variations.) For ASR, a set of speakers typically reads chosen texts, and models are developed from this speech. ASR accuracy is usually proportional to the empirical similarity between training and testing data. For example, we may get high accuracy if an ASR model is properly developed for a single speaker repeating a word many times in a quiet environment, then testing the system with new versions of that same word from that speaker in the same environment. However, if we then test on a different speaker, with a different microphone, or add some background noise, we usually get reduced (and often much lower) accuracy. This is called the mismatch problem. The challenge for ASR designers is to amass sufficient data and employ a good training algorithm. In recent years, great strides have been made toward obtaining adequate databases for training, but many speech databases are insufficiently labeled as to their content, and few are reliably labeled to a precision of individual phonemes (TIMIT is the most common one in ASR research) (see Table 1) [75]. In addition, many databases employ read speech (to facilitate the labeling process, and to allow clear scientific experiments, for purposes of control), yet practical applications for ASR involve spontaneous speech, for which ASR is much more difficult than for read speech, owing to the greater variability in speech when one has to think as one speaks (reading is simpler cognitively than spontaneous speech). In practice, models for spontaneous-speech ASR often derive from examples of conversations.

A major challenge for ASR is to overcome the ``mismatch'' problem, where very often a system is faced with testing speech that is a poor match for the speech the recognizer was trained on. Intra-speaker variability (i.e., speaker freedom) is usually handled reasonably well via statistical models. Inter-speaker variability seems to be a greater problem: in SD systems, each user trains the system to ``learn'' his voice, and only models for that speaker are examined for recognition. In SI systems, at least dozens of speakers provide multiple training tokens for each unit. The simplest approach merges all speakers into a single model for each phoneme. However, in such cases, the state probability density functions (PDFs) tend to broaden significantly (larger variances), causing reduced discrimination between unit classes. One way to reduce this loss of discrimination is to have models for different groups of speakers, e.g., for classes of speakers (e.g., men vs. women, different dialects). The disadvantage is increased computation, since the input speech passes through all potential models (running a gender or dialect detector as a precursor is rarely done, owing to the high risk of error). This approach of multiple models to handle environmental variability easily extends to background and transmission channels.

ASR systems are often speaker-adaptive: for a given input (assumed to be from a single speaker), one starts with an SI system, and then adapts the system parameters to the new individual user's voice [4–6]. (In audio-conference applications, one could also apply speaker tracking, to estimate when the input voice changes identity, so as to restart the adaptation.) Among the common methods of adaptation are maximum a posteriori (MAP) [7,8] (which often requires several minutes of training data, because only those models corresponding to identified sounds in the adaptation speech are modified), maximum likelihood linear regression (MLLR) [9,10] (which calculates transforms of speaker space using unsupervised adaptation data, grouped automatically into similar data sets), vector-field smoothing (adapting parameters across models incrementally), ``eigenvoices'' [11], and vocal-tract-length normalization (VTLN) (where one estimates a speaker's VT length [12]).

The most difficult variability that ASR must handle is due to background, channel noise, and other external distortions [13]. Basic spectral subtraction techniques can help with additive noise, while some cepstral methods (which convert multiplication in the spectral domain to cepstral addition) suppress convolutional noise. Many methods that are used to enhance noisy speech for human listening can be used as preprocessors for ASR. In noisy cases, one should focus on the high-amplitude parts of the input signal spectrum: strong speech formants are the most relevant for speech perception, and are relatively less corrupted by noise [14]. Two methods are normally used: robust parameterization (seek analysis parameters that are resistant to noise) or model transformation (adapt the ASR models to accommodate the distortion).

Cepstral mean subtraction (CMS), like RASTA processing [15], eliminates very slowly varying signal aspects (presumed to be mostly from channel distortion). The mean value for each parameter over time (typically for periods exceeding 250 ms) is subtracted from each frame's parameter, thus minimizing environmental and intra-speaker effects. Channel noise is often assumed to be constant over an utterance, but portable telephones suffer fading channel effects, which require more frequent estimations [16]. Another example of a model transformation to improve ASR is parallel model combination (PMC) [17].

1.2. Brief history of ASR

Using analog circuitry, Bell Labs demonstrated small-vocabulary recognition for digits spoken over the telephone in 1952. As