

On Integrating Insights from Human Speech Perception into Automatic Speech Recognition

Sorin Dusan and Larry R. Rabiner

Center for Advanced Information Processing
Speech and Language Processing Laboratory
Rutgers University, Piscataway, NJ 08854, U.S.A.
sdusan@caip.rutgers.edu, lrr@caip.rutgers.edu

Abstract

In spite of the effort and progress made during the last few decades, the performance of automatic speech recognition (ASR) systems still lags far behind that achieved by humans. Some researchers think that more speech data will be sufficient in order to bridge this performance gap. Others think that radical modifications to the current methods need to be made, and possible inspirations for these modifications should come from human speech perception (HSP). This paper focuses on two issues: first, it presents a comparison between HSP and ASR emphasizing some insights from HSP that could still be applied in ASR; second, it presents some ideas for extracting useful non-linguistic information from the speech signal, the so called ‘rich transcription’, which could help in selecting specialized acoustic-linguistic models that offer higher accuracy than the general models.

1. Introduction

Substantial progress has been made in the area of automatic speech recognition (ASR), especially during the last two decades, by adopting and perfecting techniques based on hidden Markov models (HMM) and artificial neural networks (ANN). Once these techniques matured and started to level off in performance, more researchers began thinking of alternative recognition architectures and principles. A key reason for examining HSP more closely is that the difference in performance between ASR and human speech recognition (HSR) is larger in tasks that involve realistic ‘noise and background’ conditions than in artificially noise-free conditions. Lippmann provides an excellent summary comparing and contrasting the performance of modern ASR systems and HSR across a range of recognition tasks [1].

The earliest attempts to perform ASR, although primitive, were based on early understanding of hearing and human speech perception (HSP). These early speech recognition systems assumed that short speech segments corresponding to phonemes, dyads, syllables or even words could be uniquely mapped into the corresponding linguistic units by measuring the “spectral distance” between these segments (or some appropriate spectral representation of these segments) and a set of previously recorded and labeled templates.

The current ASR techniques (such as HMM and ANN) are data-driven. Such data-driven models infer or learn the relevant speech structures from large quantities of training data and use relatively simple speech models to map acoustics to (context-dependent) phones and words. These techniques and the resulting models are often cited for incorporating too

little (linguistic and acoustic) knowledge about speech and about the processes of speech perception in humans. While a large number of researchers from the ASR community believe that new directions and discoveries are necessary to match human performance, more conservative researchers think that having significantly more training data will be sufficient to achieve this goal.

There are a number of theoretical and practical problems in integrating more knowledge from HSP into ASR. First, there is no complete understanding of the mechanisms and processes that take place in human speech perception and speech comprehension; second, not every complete or partial discovery in speech perception leads to an improved computational model for ASR; third, the area of speech perception and understanding is multidisciplinary where discoveries and theories emerge from fields such as neuroscience, psychology, linguistics, cognitive science, etc., and these fields are usually not well monitored and understood by the engineers and computer scientists who implement ASR technologies. Finally, due to its complexity, understanding how the brain works poses formidable difficulties even for simpler, more specialized functions such as speech perception.

Section 2 compares some general characteristics of HSP with techniques used in ASR and emphasizes some insights from HSP that have not been implemented in ASR. Section 3 provides a description of some current research directions in our laboratory for extracting a rich transcription from speech.

2. A comparison between HSP and ASR

Most of the more biologically-inspired auditory models, proposed for use in ASR systems, account primarily for the processes that take place in the cochlea and at the lower levels in the auditory pathway. These include the Lyon Cochlear model [3], the Seneff joint synchrony/mean-rate model [4], and the Hermansky Perceptual Linear Predictive (PLP) model [5]. We now make a comparison between HSP and ASR along six key dimensions of ASR.

2.1. Architecture and levels of organization

The most important architectural difference between HSP and ASR is that the former involves a large parallel neural processing system whereas the latter, based on computers, uses a serial processing system. Functionally, the former uses millions of neurons whose information processing rates are relatively low (a neuron can fire at a rate of approximately less than one thousand times per second) whereas the latter usually employs one microprocessor whose processing rate is

very high (a microprocessor can currently work at a rate of about one billion instructions per second).

Another important distinction is represented by the higher number of levels of organization in HSP than in ASR. In humans the spectral-temporal information from approximately 3,500 inner hair cells (IHCs) and 12,000 outer hair cells (OHCs) along the basilar membrane is transmitted by approximately 30,000 afferent fibers (90-95% receiving from the IHCs [7]) in each of the auditory nerves to approximately 90,000 neurons in the cochlear nucleus. Additional processing occurs at higher levels using the 34,000 neurons in the superior olivary complex and trapezoidal body, the 38,000 neurons in the lateral lemniscus, the 400,000 neurons in the inferior colliculus, the 500,000 neurons in the medial geniculate body and the 100,000,000 neurons in the auditory cortex [2]. Another hierarchical organization in HSP is represented by the six layers found in the auditory cortex. Such complex interconnections and the increasing number of carriers of information found in the auditory pathway are not usually implemented in existing ASR systems.

An important distinction between ASR and HSP comes from the existence in humans of various parallel arrangements in the thalamocortical auditory pathway, apparently specialized to transmit and process distinctive properties of the sensory information. At least three principal parallel pathways were found that correspond to a tonotopic system, a non-tonotopic system and a polysensory (multimodal) system. However, the exact contribution of these parallel systems to HSP is currently unknown. The specialization of populations of neurons was also found in the visual system where different groups of neurons process different attributes of images, such as form, color and motion [40]. In addition other groups of neurons appear to be specialized to represent more detailed characteristics of the acoustic signal. Neurons are specialized to have the highest sensitivity at a specific characteristic frequency (CF), or at a specific threshold (TH) of the sound pressure level, or to have a specific spontaneous activation rate (SA), firing range (FR), or dynamic range (DR). These types of specializations in processing the acoustic properties are not found in ASR, where the acoustic features are homogeneous, although they do represent the frequency scale in a non-linear manner.

A different architectural distinction is represented by the redundancy offered by large groups of neurons in transmitting the same or similar information to the higher levels of processing. In the cochlea, each inner hair cell transmits the spectral information to 10 or more fibers in the auditory nerve. The same type of redundancy is found at all the upper levels of processing, although the distribution and combination of information could be much higher since a typical neuron has between 1,000 and 10,000 synapses. In the brain the sounds are processed by continuously increasing the number of neurons (many carrying redundant information), whereas in ASR the sounds are parsimoniously represented by a reduced number of features (from hundreds of speech signal samples to tens of spectral features). The principle of redundancy does not play any role in current ASR techniques.

2.2. Spectral analysis and feature representation

In humans the spectral analysis is performed along the basilar membrane of the cochlea by some 3,500 IHCs and 12,000 OHCs. Although it is believed that only the IHCs

transmit 'important' ascending information to the higher auditory levels, it appears that OHCs contribute significantly to the frequency selectivity and sensitivity of IHCs. The approximate 30,000 neural fibres in each auditory nerve represent the acoustic signals by a myriad of firing rate patterns derived from all these neurons. Each of these neurons responds only to a specific frequency range and has specific characteristics (CF, SA rate, TH, DR and FR). Hence, there is a high specialization among these neurons in order to represent the entire frequency and dynamic range of human audition. In ASR the spectral analysis of the acoustic signal is performed usually at a few hundred frequency points (e.g., 512 Fourier magnitudes) which are then reduced to 10-20 spectral dimensions. Although both representations are time-dependent (time varying), in HSP the acoustic features are represented by firing rate (frequency) patterns and in ASR they are represented by magnitude patterns.

2.3. Top-down information processing

In hearing there is a large number of top-down connections represented by the efferent fibers at all the levels of the auditory system. Although their functions are far from well understood, it is believed that they play an important role in audition, including HSP. Various studies based on animals showed that the recognition of vowels in identification experiments was seriously degraded when efferent fibers in the auditory nerves were cut. In ASR such top-down connections are only implemented at high levels by emulating semantic and syntactic constraints with linguistic rules and word probabilities (n-grams). The efferent connections in the auditory pathway do not usually have any direct equivalent at the lower levels of information processing in ASR (e.g., features are not affected by top-down information). This could be a reason for the lack of robustness in ASR.

Another dissimilarity between HSP and ASR is the existence of another major level of the architecture and process in HSP, represented by the multimodal concept code level, in addition to the word code level and the syntactic-semantic level. The syntactic-semantic level influences word recognition in both systems but the multimodal concept code level is either not existent in ASR or is represented in a one-to-one manner by the word code level. This is not a minor distinction since in HSP the concept-word interaction is bi-directional and the concept code level is grounded in multiple modalities (e.g., sensory, motor, somatic). In ASR the recognized word is only influenced by the bottom-up (acoustic) and the top-down linguistic constraints (syntactic and semantic) that are also present in HSP. Individual word-concept interactions are simulated in ASR only at a linguistic level by using word probabilities (e.g., bi-grams, tri-grams, etc.) but this is a rather simplistic process and does not come from the multimodal environment context but from previous utterances (words).

2.4. Speech units

All the important speaker-independent ASR approaches use the concatenation principle to represent words by successive phonemes. The fundamental unit of processing is thus the phoneme, which is usually represented by a context-dependent model (e.g., triphone, demiphone). Words are represented in the pronunciation lexicon as concatenations of phonemes, similarly as they are represented in writing by a

concatenation of letters. However, in HSP it is unlikely that the phoneme and the concatenation principle play the only central roles as in ASR. A variety of experimental and theoretical studies provide more and more evidence that in HSP some holistic processes employing words, syllables, and transitional units (diphones) are likely to also be involved. There are very few approaches in ASR based on heterogeneous speech units.

2.5. Speech segmentation

Current ASR techniques, such as the HMM, perform the search for the best sentence by combining the acoustic and linguistic information and searching a lattice of words and phones for the best hypothesis. The recognizer exploits pauses or silence intervals in the incoming speech for segmenting the utterance into sentences (or phrases). Because these sentences usually obey grammatical rules, these rules can be imposed by top-down syntactic-semantic constraints, and thus the recognizer performs better when the recognition and final segmentation are performed on the whole sentence. The segmentation into words precedes the recognition and there is usually no local effect of the recognition of the current word on the segmentation of the next word.

In HSP it appears that the perception of the 'current' word plays an important role on the identification of the onset of the following word (segmentation), whose onset identification in turn, plays a very important role on the recognition of that word. This might be also influenced by the fact that in HSP the meaning of an individual word usually is perceived at runtime ('instantly') and not after the completion of the whole sentence, although there are situations when that happens (sometime the recognition of the meaning of a current word depends upon a word or phrase that only comes after a few more words). It should be understood that the 'current' word is processed with a certain delay and that this delay may not be always constant and may depend upon the identity of the word and the context in the sentence. In reality, some words (in particular long words) appear to be recognized before they end. Such a local effect imposed by the recognition of the current word upon the segmentation of the following word is not directly implemented in ASR where multiple segmentation hypotheses are derived by imposing semantic and syntactic constraints and precede the final recognition of the whole sentence. However, in principle, such an approach could be implemented in ASR by providing onset markers after recognition of high-confidence words and associating additional probabilities with these segmentation markers based on the confidence measure of the preceding word.

2.6. Speech variability

One of the most difficult problems in ASR is dealing with the great variability found in natural speech and with the effects of various noisy environments. Humans perform much better than ASR in perceiving speech from many speakers and different environments without a prior exposure to the exact type of speech and environmental noise. That leads us to suspect a deficiency in ASR in coping with the large variability in speech and environmental factors.

Speech rate variability has a wide range. While a normal speech rate comprises about 10 phonemes per second, speech can be produced and perceived at much lower and higher rates. For example, the comedienne Fran Capo, who is listed

in The Guinness Book of World Records as the fastest talking female, has achieved an incredible speaking rate of 603.32 words per minute, which is about 10 words per second. It is likely that the brain deals with the high degrees of spectral and temporal variability of speech by employing specialized neural mechanisms.

In HMM, temporal variability is modeled by transitional probabilities (or by explicit state or phone duration models) among stationary (or non-stationary) phone states, and spectral variability is usually accounted for by employing a mixture of a large number of Gaussian densities to represent the composite probability density of the acoustic features of the phones in each state. In humans, as the information proceeds from the cochlea to higher hierarchical levels in the auditory pathway, an increase in the dimensionality and the heterogeneity of this space takes place (30,000 in the auditory nerve and 500,000 in the medial geniculate body before entering the auditory cortex). Heterogeneous spaces of representation (increased by parallel neural channels specialized for ranges of variability) could explain in part the high performance of HSP in adverse conditions (e.g., new speakers and environments). Hence, in general, in HSP the variability of speech is more likely accounted for by many parallel sets of neurons that map into the same higher level representation of a speech segment (phoneme or phonological sequence), each specialized to cover a specific region of variability. The distinction is that these speech segments might be represented by heterogeneous multi-dimensional spaces, i.e., using different heterogeneous spaces specialized for particular speech segments and particular ranges of variability.

3. Rich transcription

Certainly humans extract a wealth of non-linguistic information from speech that plays an important role in speech perception and understanding. In fact this useful information in speech understanding is also complemented by the use of other input modalities (vision, touch, smell, etc.). In ASR the usual task is word transcription, which is somehow artificial because it is isolated from a variety of other cognitive processes that normally take place in the human brain. Extracting non-linguistic information from the speech signal can be viewed as a method of preprocessing. Some of the attributes of the resulting rich transcription, include speaker's characteristics (e.g., gender, height, weight, vocal tract length, accent, emotion, etc.), speech non-speech distinctions, segregation of multiple/overlapped speech streams, sense of distance to the speaker, and awareness of the use of an unknown foreign language.

Current experiments are underway in our laboratory for estimating some of these non-linguistic features. A method for speech/non-speech detection based on recognizing transitional (diphone) units in audio streams is being studied. A new method of detecting overlapped speech by building Gaussian mixture models is also under investigation. We are looking at methods for speaker segmentation of multi-speaker streams, and finally we are studying a new method for estimating speaker's height, vocal tract length, weight, and gender. Some of these results regarding the estimation of speaker's height and vocal tract length from speech are presented separately in another paper [8].

4. Discussion

Significant progress has been made in understanding brain and language processes during the last two decades [9]. The current paper points to some potential sources of new ideas, based on this research progress, for improving ASR performance.

There is increasing evidence that transitional units play a very important role in HSP, maybe even more important than that of phonemes. It might also be possible that words are perceived, both holistically and microscopically (possibly involving various phonological units or features), and the information emerging from various levels is integrated leading to a decision among competing candidates. It is known that humans can accurately perceive isolated words and syllables (or even some phonemes) that can be produced in isolation. That simply means that humans have the ability to perceive various speech 'codes' even without a larger context or without meaning. This does not mean that the brain uses exactly the same processes and channels in all these tasks or that the accuracy of identification is the same in all these cases. However, the alternative suggested here for speech perception is different from that suggested in [6], which supports the idea that there is no fundamental unit in speech perception and the perceived unit is the one that attention is focused on in a specific task. The idea suggested in this paper agrees that in the perception of words the objects of attention are the words, whereas in the recognition of nonsense syllables or phones the objects of attention are these units. But, in addition, this paper suggests that the perception of words involves a few parallel processes that all concur in the perception of the word. That is, there are simultaneously involved processes: a holistic process of word perception and a few other sub-word processes such as for the recognition of syllables, diphones, and certain phones.

At the phoneme level the information specifying the phoneme category is distributed across a continuous time interval that extends beyond the currently considered phoneme boundaries in ASR. An analysis of this multi-level information based on seven distinctive acoustic cues in the identification of vowels is presented in [10].

The most important argument supporting the multi-level model of word perception is that all of these units are repeatedly heard during the process of language acquisition, and this must inevitably lead to the creation of architectural patterns of synaptic connectivity in the auditory pathway and the auditory cortex at various hierarchical levels and not only to a single pattern ending with the word 'code'. It is known that such synaptic connections and the early beginnings of recognition of words occur during the first year of infancy whereas the acquisition of the meaning of the words only begins during the second year of the child's life. Since the absence of the meaning of words does not preclude the building of such word codes from continuous speech, it is unlikely that the brain does not build similar codes for syllables, diphones or even phonemes due to a lack of meaning when they are heard repeatedly during language acquisition.

Perceptual studies on the spectral transitions between phonemes show evidence that these regions play a very important role in speech perception [11]. If these transitions are so important then why should not the brain have individual recognition 'codes' for them, since they are not

characteristics of individual phonemes (they do not belong to phonemes) and they only characterize specific combinations between phonemes. Since humans usually retain a lexicon comprising a few tens of thousand of words, what would be the memory economy of not employing a few more thousand codes for these important phonological segments?

The multi-level model of word perception is well supported by principles of redundancies, which are considered to play an important role in HSP and perception in general. It is also supported by the availability of an immense number of neurons in the auditory sensory system. It is supported by the evidence that the brain has segregated (specialized) areas for the processing of various characteristics of the sensory information (e.g., in vision - form, color and movement). The model proposed here is merely a sum of ideas instead of a theory. Preliminary experiments implementing these ideas are newly underway.

5. Conclusions

It appears that existing ASR technologies need radical modifications in order to bridge the performance gap between HSR and ASR. For potential new directions in ASR, a good place to start is HSP, although it might be possible that some new mathematical approaches would provide better solutions than currently achieved by ASR. It is argued here that to bridge the performance gap new ASR systems also need to extract more non-linguistic information from the speech signal.

6. References

- [1] Lippmann, R. P., "Speech recognition by machines and humans," *Speech Communication*, 22: 1-15, 1997.
- [2] Worden, F. G., "Hearing and the neural detection of acoustic patterns," *Behavioral Science*, 16: 20-30, 1971.
- [3] Lyon, R. F., "A Computational Model of Filtering, Detection, and Compression in the Cochlea, Proc. IEEE-ICASSP-82, 1282-1285, 1982.
- [4] Seneff, S., "A joint synchrony/mean-rate model of auditory speech processing," *Journal of Phonetics*, 16 (1): 55-76, 1988.
- [5] Hermansky, H., "Perceptual Linear Predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, 87(4): 1738-1752, 1990.
- [6] Goldinger, S., and Azuma, T., "Puzzle-solving science: the quixotic quest for units in speech perception," *Journal of Phonetics*. 31: 305-320, 2003.
- [7] Rouiller, E., "Functional Organization of the Auditory Pathways," *The Central Auditory System*, New York, Oxford, Oxford University Press, pp. 3-96, 1997.
- [8] Dusan, S., "Estimation of Speaker's Height and Vocal Tract Length from Speech Signal," submitted to EUROSPEECH 2005.
- [9] Damasio, A. R., "Brain and language: what a difference a decade makes," *Current Opinion in Neurology*, 10: 177-178, 1997.
- [10] Dusan, S., "On the nature of acoustic information in identification of coarticulated vowels," submitted to EUROSPEECH 2005.
- [11] Furui, S., "On the role of spectral transition for speech perception," *J. Acoust. Soc. Amer.*, 80(4): 1016-1025, 1986.