Overview of Auditory Strategies for Speech Recognition

A review of some related papers and proposed research topics

Alex Park
malex@sls.lcs.mit.edu

June 6, 2003

1 Introduction

The purpose of this report is to outline several directions for research in the general area of robustness in recognition, and to review related work in these areas. The running theme in each of these areas is the idea of borrowing strategies from the human auditory system to improve speech discrimination in environments where traditional ASR approaches fail. Since it is more likely that speech has adapted to complement the strengths of the auditory system and not the other way around, it is also likely that we can learn more about speech by investigating properties of the auditory system.

Some speech scientists contend that copying the behaviour of the auditory periphery is simply mimicry and will therefore not lead to an understanding of the underlying principles behind speech processing. Indeed, many phenomena that occur in the auditory pathway may be completely irrelevant to the task of speech recognition. In [1], Hermansky contrasts modern airplanes with birds to illustrate that copying nature is not always an optimal way to achieve some desired objective. He remarks that there is an important difference between understanding the principle behind a process and just copying the outward appearance of the process. We agree with this viewpoint, and seek to study the auditory system not for its own sake, but to discover useful principles that can be applied to automatic systems. It should be noted that important properties of the auditory system have worked their way into the front end processing of most ASR systems. For example, computation of Mel-scale cepstral coefficients includes frequency analysis (FFT), non-uniform frequency resolution (Mel-scale filtering), and dynamic range compression (logarithm), all of which are performed in the auditory periphery.

A second argument that is often raised regarding auditory strategies is that they represent a step backward from current ASR systems which are able to at least generate a recognition result. Indeed, while auditory strategies attempt

to make use of measurements that may be more like the cues humans use to decode speech, these cues cannot be easily incorporated into existing ASR systems. Although the end goal of recognition is important, we believe that the current complexity of existing systems, with their abundance of parameters, makes it difficult to understand what factors are contributing to improvements in recognition accuracy.

2 Cochlear non-linearity

A great deal of hearing research is focused on trying to understand the causes and effects of cochlear non-linearity. This phenomenon refers to the fact that the local frequency analysis performed along the basilar membrane in the cochlea is compressive, with level dependent filter characteristics. For any particular characteristic frequency (CF) along the membrane, the magnitude response of the filter tuned to that particular frequency has sharper tuning to the CF for low level inputs, and broader tuning at higher level inputs.

Although it is known that the nonlinear behaviour of the basilar membrane is caused by the outer hair cells, the mechanisms behind this process are not well understood. While we may not know how this is accomplished in the ear, the net effects of the cochlear nonlinearity are at least twofold. First, the compression of dynamic range that occurs is not uniform across all frequency channels (Differential compression). Second, frequency channels are not laterally independent, since an off-frequency input can inhibit the response of a filter tuned to a particular frequency (multitone suppression). The question is, are these effects important for speech processing?

A compelling argument for including the effects of cochlear nonlinearity in the front end for ASR systems is provided by speech reception experiments on hearing impaired individuals. There is a considerable body of work which indicates that people with damaged outer hair cells tend to have good speech recognition ability in quiet environments, but this ability degrades much more drastically than normal hearing listeners in noisy environments [2]. It is possible that the absence of a compressive cochlear non-linear component, which is common to automatic systems and hearing impaired listeners, may explain poor speech discriminability experienced by both in noisy environments.

2.1 Recent Work

There are two recent computational approaches that have been proposed to model the nonlinearity at the level of the basilar membrane.

2.1.1 Gammachirp filterbank

This nonlinear filterbank, proposed by Irino and Patterson [3], [4], is a variation of the popular gammatone filterbank which includes a chirp parameter in the impulse response of each filter. In the frequency domain, gammachirp is shown

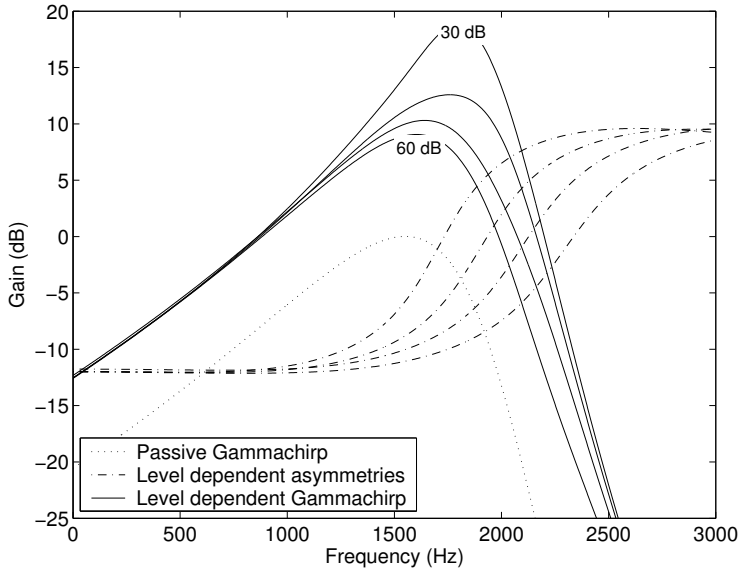

Figure 1: *Gammachirp filter.*

to be a cascade of the symmetric gammatone filter with an asymmetric component. By making the symmetric component dependent on the level of the input stimulus, the peak gain and tuning characteristics of the individual filters can be made to match human masking data. An example of the filter characteristics for several different input levels are shown in Figure 1. The disadvantage of the gammachirp filter is that it requires a parameter controller to modify the characteristics of the asymmetric component based on the input level. To that end, it is necessary to keep a running measure of the current input level for each frequency channel and update the filter parameters at some time interval.

This filter has been used in [5] as a method for speech enhancement by resynthesizing the output of the gammachirp filter. This approach was shown to be good at suppressing broadband stationary noise without introducing the "musical" distortion that is associated with spectral subtraction.

2.1.2 Dual Resonance Non-linear (DRNL) filterbank

The DRNL filterbank has been proposed by Meddis et. al as an alternative to the quasi-nonlinear gammachirp filter [6], [7]. This approach uses two parallel pathways to construct the overall output response to an input. The first pathway is a broadly tuned gammatone filter which is meant to simulate the basilar membrane response at high levels. The second pathway is the nonlinear pathway, and is a cascade of narrowly tuned gammatone filters which are sandwiched around a nonlinear gain function. The nonlinear gain amplifies lower level inputs more than higher level inputs.

The output from each of the two pathways are summed together to get the final output. At low levels, the nonlinear gain function in the narrowly tuned pathway causes the narrow gammatone to be the dominant component in the overall output. At high levels, the nonlinear gain function attenuates the output in the narrowly tuned pathway, which causes the broadly tuned gammatone filter to be the dominant component. The overall effect is a level-dependent filter which has been shown to exhibit

2.2 Questions

The major research questions involved in this stage are the following:

- How can we characterize the differences between the outputs of linear and nonlinear auditory filterbanks?
- Are these differences actually useful for speech processing in noise? If so, how can we demonstrate their utility, and make use of them?

3 Speech-relevant features

Another direction for research is to identify features and cues that humans use for speech processing such as pitch, voicing, formant motions, amplitude modulation, and temporal information. More importantly, it is necessary to propose reliable methods for computing such features.

The rationale behind discarding pitch information for MFCCs is to improve generalization among different speakers. However, the fact that MFCCs are equally useful for speech recognition and speaker identification indicates that MFCCs are probably not an optimal feature set for either task.

3.1 Related Work

Seneff [8] proposed a model of the auditory periphery for initial processing of sound and followed that stage with two pathways. The first path calculated mean firing rates of auditory nerve fibres, and was a measure of spectral energy. The second path utilized a novel technique for detecting synchrony of fine temporal structure in each channel. This technique was extended to provide a mechanism for pitch detection and estimation based on temporal information.

Ali extended the synchrony detector in the Seneff auditory model to include lateral synchrony detection (repetition cues from adjacent channels) [9]. He made use of this measure to determine acoustic-phonetic features for discriminating between stop consonants [10], and fricatives [11]. Although these features were tested using only phoneme classification experiments, he proposed that such a rule-based classification scheme could be used to supplement traditional statistical model-based systems.

Saul et. al proposed the use of voicing as a cue that can be robustly extracted in adverse acoustic environments. They used an auditory front end

followed by periodicity and SNR measurements on each channel. These narrowband measurements were then combined to produce a global decision using a structured multilayer Bayesian network. The rationale for not using fullband measurement directly was that training classifiers on fullband data only would cause the classifier to rely on portions of the frequency spectrum that are only reliable in clean speech. The narrowband approach essentially trades away some performance on clean speech for generalization ability in noise. In noise, they were able to show significantly lower detection error than a traditional Gaussian classfier trained on MFCC features.

3.2 Questions

Some research questions involved in this stage are the following:

- What are some relevant acoustic features that humans use when decoding speech?
- How can we extract such features in a manner which is reliable across noisy conditions?

4 Flexible Recognition using Multiple Cues

A final direction for research proposed here concerns the integration of diverse features and cues to generate a recognition hypothesis. All of the works described in the previous section fell short of actually producing a full speech recognizer. In order to perform recognition, lower level features must combined in an appropriate way to give information about the type of speech perceived. An ideal integration strategy must be able make use of highly discriminative features in clean speech, but should also be able to back off to less discriminative, but more reliable features when dealing with noisy or corrupted speech. Moreover, higher level top-down knowledge should be incorporated into the recognition process when it is available, such as visual input, knowledge of current domain, location, etc.

Current ASR systems are inflexible from the bottom up in that they are trained on a single set of features and cannot cope with any more or any fewer than those features. They are also inflexible from the top down because they generally constrain the search space to be either speech or non-speech from a single speaker. These assumptions ensure better performance in domains where the input is constrained, but drastically limits the generalization ability of the systems built on this basis. In particular, the inability to detect OOV words and triggering of word hypothesis for non-stationary, non-speech sounds, is a conspicuous result of these constraints.

4.1 Related Work

Cooke and Ellis have proposed a method for improving flexibility in recognition by allowing for missing features in the lower level representation, and by search-

ing over likely speech fragments in the higher level stages [12]. The output of an auditory filterbank is used as the feature set, and an estimate of the SNR is used to generate a reliability mask of time-frequency fragments determined to be more reliable than others. These fragments are then split again and the marginalisation is used to determine likelihood scores for frames with missing features. The final decision to include a spectral fragment is made in the search stage, which allows higher level knowledge to be incorporated in this decision.

4.2 Questions

Some research questions involved in this stage are the following:

- What kind of framework can we use to combine time-varying features, and how can we make such a framework flexible enough to deal with subsets of those features.
- What is an intelligent way to incorporate higher-level knowledge to determine the reliability of lower-level features.
- What type of lexical representation is needed to support a non-frame based feature set?
- How can we handle the problem of search when the search space is not tightly constrained?

5 Conclusion

The questions posed in this report represent potential directions for future research. *To be completed.*

References

[1] H. Hermansky, "Should recognizers have ears?," *Speech Communications*, vol. 25, no. 1-3, pp. 3–27, 1998.

[2] R. Peters, B. C. J. Moore, and T. Baer, "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," *J. Acoust. Soc. Am.*, vol. 103, no. 1, pp. 577–587, January 1998.

[3] T. Irino and R. D. Patterson, "A time-domain, level-dependent auditory fiilter: the gammachirp," *J. Acoust. Soc. Am.*, vol. 101, no. 1, pp. 412–419, January 1997.

[4] T. Irino and R. D. Patterson, "A compressive gammachirp auditory filter for both physiological and psychophysical data," *J. Acoust. Soc. Am.*, vol. 109, no. 5, pp. 2008–2022, May 2001.

[5] T. Irino, "Noise suppression using a time-varying, analysis/synthesis gammachirp filterbank," in *Proc. ICASSP*, Phoenix, AZ, 1999.

[6] R. Meddis, L. O'Mard, and E. Lopez-Poveda, "A computational algorithm for computing nonlinear auditory frequency selectivity," *J. Acoust. Soc. Am.*, vol. 109, no. 6, pp. 2852–2861, June 2001.

[7] E. Lopez-Poveda and R. Meddis, "A human nonlinear cochlear filterbank," *J. Acoust. Soc. Am.*, vol. 110, no. 6, pp. 3107–3118, December 2001.

[8] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *Journal of Phonetics*, vol. 16, pp. 55–76, 1988.

[9] A. M. A. Ali, J. Van der Spiegel, and P. Mueller, "Robust auditory-based speech processing using the average localized syncrony detection," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 279–291, July 2002.

[10] A. M. A. Ali, J. Van der Spiegel, and P. Mueller, "Acoustic-phonetic features for the automatic classification of stop consonants," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 833–841, November 2001.

[11] A. M. A. Ali, J. Van der Spiegel, and P. Mueller, "Acoustic-phonetic features for the automatic classification of fricatives," *J. Acoust. Soc. Am.*, vol. 109, no. 5, pp. 2217–2235, May 2001.

[12] M. Cooke and D. P. W. Ellis, "The auditory organization of speech and other sources in listeners and computational models," *Speech Communications*, vol. 35, no. 3-4, pp. 141–177, 2001.