BETTER ANALYSIS FOR AUTOMATIC SPEECH RECOGNITION

Douglas O'Shaughnessy
INRS-Telecommunications (EMT, University of Quebec)
800 de la Gauchetiere west, Montreal, Quebec, Canada H5A 1K6
dougo@inrs-telecom.uquebec.ca

1. INTRODUCTION

Automatic speech recognition (ASR) of speech appears at first glance to be a simple task. Commercial systems often claim to do ASR reliably, but they usually require high-quality voice input and often impose many restrictions on what is said and how speakers talk. In addition, recognition of speech in poor acoustic conditions is often unreliable. Another consideration is that of computer resources. When ASR is done at a central computer, where system speed and memory is less of a concern, this latter issue may not be so important, but when ASR occurs in portable devices with limited power and memory, minimization of resources becomes important. We discuss in this paper an efficient ASR analysis method, which applies to adverse acoustical conditions.

2. BACKGROUND

A major problem for most ASR systems is robustness: they often are insufficiently general or are over-trained when furnished with small training sets (as typically happens in many practical cases). An ideal robust ASR system should be able to properly decode speech from all speakers of a language (e.g., English), in any reasonable environment, and with different microphones and transmission channels. In practice, instead, environmental noise (from natural sources or from machines) and communication distortions in transmission channels (e.g., static, fading) both tend to degrade ASR performance, often severely. Human listeners, by contrast, usually can adapt rapidly and successfully to most such difficulties. This large difference between machine and human speech recognition performance strongly suggests that major flaws exist in current ASR schemes. In particular, much of what the scientific community knows about human speech production and perception has yet to be properly integrated into practical computational ASR.

The speech signal must be regularly converted to a representative, small set of parameters or features, in order to efficiently and reliably interpret the audio signal (input to ASR) in terms of phonemes and words. The most common analysis method for today's ASR is the mel-frequency cepstral coefficient (MFCC) approach [1]. In a first step, either an FFT (fast Fourier transform) or LPC (linear predictive coding) spectrum is obtained using each speech frame as successive input. Then, for each frame, the logarithm of the amplitude spectrum is taken (converting to the decibel scale). Thirdly, a set of about 20 triangular filters, spaced according to the perceptual mel (or bark) scale, weights this result, yielding a simple set of 20 output energies. Finally an inverse FFT using the 20 energies as input is performed [2]. The low-order coefficients (e.g., 10-16 in number) of this last step provide the spectral vector for ASR use.

Among the advantages of this standard approach are the following: 1) an automatic and efficient method, which needs no controversial (i.e., error-risking) decisions, 2) actual ASR results that appear to be better than with some other methods extensively examined in the past (e.g., basic LPC, or a filter bank), and 3) an elegant mathematical interpretation of the MFCCs as somehow decorrelated (because the inverse FFT uses orthogonal sinusoidal basis functions). Despite their considerable popularity, however, the MFCCs are suboptimal:

1) The fourth step of the MFCC calculation (the inverse FFT - effectively low-order cosine weightings of the log spectral weighted energies) is motivated almost entirely on mathematical grounds, rather than communicational or scientific reasoning, which has led to representing spectral information for speech in a very convoluted way. The first output coefficient (C0, which uses a zero-frequency cosine weight) is simple energy, hence easy to interpret and utilize (if desired). The second (C1, using a cosine whose spectral period comprises the full frequency range) is thus a spectral parameter which indicates the global energy balance between low and high frequencies (the initial, positive half of the cosine weights the lower half of the frequency range positively, and vice versa for the upper range). Thus, the first two coefficients are useful and subject to easy interpretation. However, all the other MFCCs are very difficult to relate to major aspects of speech production or perception. For ASR purposes, it is not essential that the parameters used be physically interpretable, but the MFCCs must be used altogether, in order to exploit the fact that they contain increasingly finer spectral detail (as the order increases). Although most individual MFCCs have little clear meaning (in terms of acoustics, the vocal tract shape, or phonemes), used together they can discriminate different sounds. Unfortunately, their lack of direct or simple correspondence to speech production and perception means that they are very vulnerable to degradation when speech occurs under non-ideal acoustic conditions, such as in noise or with speakers having foreign accents.

2) It has often been posited that the MFCCs are uncorrelated in some sense, owing to the orthogonal basis functions used in the inverse FFT [2]. It is quite evident, however, that the MFCCs contain much overlapping spectral information, which causes the covariance matrices of their joint probability densities, as used in ASR, to be far from diagonal. This in turn leads to poor modeling assumptions in many ASR

applications which do indeed often assume diagonal matrices (to cut computational costs), or to significantly increased computation to handle large general matrices [2] (for the minority of cases that use full-covariance matrices). As model order for MFCCs increases, of course, these matrices grow in size proportionately. Requiring 10-16 parameters (or more) becomes increasingly expensive in storage, computation, and training. The correlation of MFCCs is easily seen in the example that both C0 and C1 share large positive values for vowels and are negative for fricative phonemes.

3) Different speakers (especially those with different accents) exhibit varying spectral patterns when uttering (what listeners interpret to be) the same phoneme. Many of the variations often have simple interpretations acoustically and linguistically; hence adaptation to such variation should not theoretically present such a large problem. In practice, the ASR field spends much research on adaptation issues, both those due to varying speakers and to varying transmission channels. If we employ spectral parameters that have a ready physical interpretation, it is feasible to model accent and channel variation simply. However, the lack of such ability to interpret the MFCCs usually forces ASR to employ "brute-force" methods, e.g., simple averaging of distributions to handle different speakers and channels. Such merging of data models often leads to increased variances and hence to lowered discriminability against other, incorrect phoneme models. For these and other reasons, the MFCCs should not be considered as the ideal speech analysis tool, despite their recent popularity.

3. ALTERNATIVE SPECTRAL MEASURES

In the early stages of serious ASR work, formant frequencies were considered the primary objectives of speech analysis. Expert system approaches to ASR abounded then, and formants were widely accepted as the obvious targets for speech analysis. Unfortunately, the automatic formant estimation methods of the 1970s (needed for "automatic" speech recognition) failed to achieve sufficient accuracy. Formants were difficult to follow reliably, as they often approached each other close enough to be viewed as a merging (in spectral displays, such as the FFT) and the formants varied widely in amplitude as a function of time.

We do not propose yet another attempt at formal formant trackers, for two reasons: 1) the continued difficulty of formant tracking, and 2) formants (as such) are not required for ASR. In our opinion, it was an error for ASR researchers to insist on a strict formant tracker as a separate module for ASR. Indeed, robust spectral measures better than MFCCs are quite feasible based on spectral peaks similar to formants, and this is where we propose to raise ASR accuracy. When faced with increasingly noisy speech (as is found in many practical ASR conditions), the peaks of such speech spectra are the most robust (i.e., the last aspects to be lost as noise grows). More robust ASR is thus possible by directly exploiting peaks, rather than approaches that deteriorate quickly in noise (e.g., MFCCs or LPC).

Trying to consistently track all the formants was a mistaken and unnecessary task for ASR. Instead, identifying the major spectral peaks in an utterance and their gross temporal dynamics are what appears to be crucial for ASR. In other words, coarse detail about spectral peaks is important; precise tracking of formants is not. We do not need formants identified as F1, F2 and F3 for all speech frames. Instead, we propose a spectral-peak-based analysis measure which can be simultaneously robust, informative, and efficient. Such a measure needs as few as six coefficients to represent the main spectral peaks (three center frequencies and their coarsely-measured bandwidths), and thus is clearly more efficient for ASR than either LPC or the MFCCs.

For noisy telephone digit strings, our method can achieve good recognition rates, without requiring the complexity of full mel-cepstral evaluation and avoiding the large search calculations of a full HMM approach. As noise levels are increased, the weaker portions of the telphone-band spectrum are increasingly obscured, but sufficient information remains concerning the spectral peak positions of the lower formants to allow digit discrimination, even in significant noise. Mistakes confusing 5 and 9 are common when the noise obscures most of the consonant energy in those digits, although the coarticulatory effects of the consonants (labial in 5 and alveolar in 9) permit some discrimination even when the consonants are fully obscured. Allowing a comparison focussed on critical frames at the ends of the vowel (rather than a uniform frame-based method) permits better utilization of the speech energy in the presence of noise. More details of the results will be presented at the conference.

4. CONCLUSION

A case can be made that the current HMM-MFCC approach to ASR has sufficient flaws as to need eventual replacement. Certainly the persistence of high error rates for many tasks that humans find easy argues that incremental improvements may well not be enough to render current ASR suitable for widespread applications. The ASR of the future must be both knowledge- and stochastic-driven.

5. REFERENCES

1) Davis, S. & Mermelstein, P. (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. SP, 28, 357--366.

2) Rabiner, L. and Juang, B. (1993) Fundamentals of Speech Recognition (Prentice-Hall: Englewood Cliffs, NJ).