Recognition of digit strings in noisy speech with limited resources

Douglas O'Shaughnessy and Marcel Gabrea

INRS-Telecommunications
900 de la Gauchetiere west, Montreal, Quebec, Canada H5A 1C6
dougo@inrs-telecom.quebec.ca

ABSTRACT

Automatic recognition of continuously-spoken digits (e.g., telephone numbers or credit card numbers) is feasible with excellent accuracy, even for speaker-independent applications over telephone lines. However, even such relatively simple recognition tasks suffer decreased performance in adverse conditions, such as significant background noise or fading on portable telephone channels. If an application further imposes significant limitations on the computing resources for the recognition task, then robust limited-resource speech recognition remains a suitable challenge, even for a vocabulary as simple as the digits. Since connected-digit recognition over telephone lines is a very practical application, the amount of computer resources needed for a given level of recognition accuracy was investigated for different acoustic noise conditions. Rather than use a traditional hidden Markov model approach with cepstral analysis, which is computationally intensive and does not always work well under adverse acoustic conditions, simpler spectral analysis was used, combined with a segmental approach. The limited nature of the vocabulary (i.e., 10 digits) allows this simpler approach. High recognition accuracy can be maintained despite a large decrease (vs. traditional methods) in both memory and computation.

1. INTRODUCTION

Automatic speech recognition (ASR) of spoken digits appears to be a simple task. Commercial systems do it reliably in many cases, but require reasonably high-quality voice input. Recognition in poor acoustic conditions, on the other hand, is often unreliable. A secondary, but important, consideration is the use of computer resources. When the ASR is done at a central location, where system speed and memory is less of a concern, the issue is perhaps not as important, but when the ASR is attempted in portable devices with limited power and memory, minimization of resources is a very practical concern. We introduce an efficient ASR method for digit recognition, which can apply to other small-vocabulary ASR tasks and which can work in adverse aural conditions.

2. BACKGROUND

A major problem for most ASR systems is robustness: often they are insufficiently general or are over-trained when furnished with small training sets (as occurs in many practical cases). An ideal robust ASR system should be able to properly decode speech from any speaker of a chosen language (English in our case), in any reasonable environment, and with different microphones and transmission channels. In practice, environmental noise (from natural sources or machines) and communication link distortions (e.g., static, fading) both tend to degrade ASR performance, often severely. Human listeners, by contrast, usually can adapt rapidly to such difficulties, which strongly suggests the existence of major flaws in current recognition schemes. In particular, much of what we know about human speech production and perception has yet to be properly integrated into practical ASR.

For various reasons (explored further below), certain procedures have become de facto standards in the ASR field: 1) the incoming acoustical speech signal is divided into frames updated every 10 msec, 2) spectral analysis reducing the input frame to a vector of 10-50 parameters is obtained, and 3) hidden Markov models (HMMs) are used to decode the speech. We will examine these, pointing out potential flaws and suggesting alternative solutions.

3. FRAME DIVISION

Since speech is dynamic (i.e., the vocal tract is constantly changing shape to communicate the speech sound sequence), one of the following is needed: a) repeated (perhaps periodic) analysis of limited window size, or b) a much broader, global time-frequency analysis. For simplicity, the former is standard, but the common choice of a fixed window length is likely not appropriate for all sounds ranging from abrupt stop closures to lengthy vowels. The 10-msec update rate (and related 25-50-msec overlapping window length) is a compromise chosen for uniformity, and to accommodate the fixed-frame-based HMM method. For the purposes of this paper (and in the interests of minimizing computer resources), we retain this approach, but we also note that better ASR results may well be possible by varying window size (and related spectral resolution); e.g., long, steady vowels may benefit from a more precise frequency resolution than a short window permits, and stop explosions may be better analyzed without smearing with nearby sounds.

4. SPECTRAL ANALYSIS

The most common analysis method for ASR is the mel-frequency cepstral coefficient (MFCC) approach [1]. Either an FFT (fast Fourier transform) or LPC (linear predictive coding) spectrum is obtained using each speech frame as input, for which the logarithm is then taken (converting to the decibel scale), a set of about 20 triangular filters spaced according to the perceptual mel scale weights this result, and finally an inverse FFT on the 20 energies is done [2]. The low-order coefficients (e.g., 10-16 in number) of this last step provide the spectral vector. Among the advantages of this approach are the following: 1) an automatic method needing no controversial (i.e., risking
error) decisions, 2) ASR results that appear better than with some other methods (e.g., simple LPC, or filter bank), and 3) an interpretation of the MFCCs as roughly decorrelated (since the inverse FFT uses orthogonal sinusoidal basis functions).

Despite their popularity, MFCCs are suboptimal:

1) the final step of the MFCC calculation (inverse FFT - effectively low-order cosine weightings of the log spectral weighted energies) is arbitrary and motivated almost entirely on mathematical grounds, encoding speech spectral information in a convoluted way. For example, the first output coefficient (C0) is energy and the second (C1) indicates the global energy balance between low and high frequencies, but all the other MFCCs are very difficult to relate to aspects of speech production or perception. They must be used in concert to exploit the fact that they contain increasingly finer spectral detail (as the order increases), which altogether allow discrimination between similar sounds. Their lack of correlation with clear aspects of speech production and perception leave them highly vulnerable to non-ideal acoustic conditions such as noise or accents. In particular, each MFCC is affected by all frequency ranges. Since most acoustic distortion is not constant across frequency (e.g., white noise is atypical), merging all frequencies together limits the ability to resist frequency-specific noise. When speech is distorted, the MFCCs are affected in widely varying degrees and complex fashion, depending upon the nature and level of the noise.

2) The spectral precision of the MFCCs is directly related to their number, e.g., for a speech bandwidth of 4 kHz (typical for telephone applications) and 10 coefficients, the last MFCC uses a cosine weight with period of 400 Hz, thus discriminating no better than an average of 200 Hz (being more than 10 cents could raise precision, but at ever increasing cost). The choice of 20 critical-band weighting filters also limits precision. While such CBs are reasonable auditory models, reducing a speech frame’s output to 20 energies is much more simplistic than the information leaving the ear on the auditory nerve. Difference-limn experiments on formants have suggested human perceptual precision as low as a few percent. Our proposed replacement for the MFCCs is not as limited in spectral precision as the MFCCs are, and the spectral precision will not vary with the number of parameters.

3) The MFCCs are purported to be uncorrelated, due to the orthogonal functions of the inverse FFT. They nonetheless clearly contain overlapping spectral information, which makes the covariance matrices of their joint probability densities far from diagonal. This in turn leads to poor modeling assumptions in many ASR applications that assume diagonal matrices (for cost-efficiency), or to significantly increased computation to handle general matrices [2] (for the minority of cases that use full-covariance matrices).

4) When different speakers (especially with different accents) exhibit similar spectral patterns for the same phoneme, the lack of interpretability of the MFCCs forces ASR to use simple merging of distributions to handle different speakers. Such merging leads to larger variances and hence lowered discriminability against other phoneme models. One recent experiment showed quantitatively the weakness of MFCCs for ASR; it asked listeners to interpret speech as passed through an MFCC processor, and found significant loss of accuracy [5].

Few researchers have directly challenged the MFCCs. Even the very few cases in the recent literature involving possible alternatives do not describe the flaws of MFCCs. For example, in [4], a fairly complex procedure yields a modest accuracy increase (over MFCCs) for ASR at 10 dB SNR, but the reasons for the improvement are not explained. In [5], the authors indeed explore relevant aspects of spectral measures, but still use the cepstral approach nonetheless. Another recent work on speaker identification [6] finds MFCCs lacking as a parameter set, but takes quite a different approach from the spectral peak-based method we propose. Very little research has actually been done on the value of MFCCs (for ASR) versus other alternatives (the original work [1] was hardly definitive for ASR).

5. ALTERNATIVE SPECTRAL MEASURES

In the early stages of serious ASR work (i.e., the late 1960s and early 1970s), formant frequencies were considered the primary objectives of speech analysis. Speech production and perception research had shown a clear correlation between the positions of the lowest three formant center-frequencies (F1, F2, F3) and vocal tract configurations, and hence with phonemes. The relationship was complex but direct, and had the advantage of transparency and interpretability. Unfortunately, the automatic formant estimation methods of the 1970s failed to achieve sufficient ASR accuracy. Frequently, formants were difficult to track reliably, as they approached each other at times and varied widely in amplitude. For an adult male vocal tract, it was known that the average spacing between formants was 1 kHz, but the range for F1 (about 300-500 Hz) overlapped somewhat with that of F2 (about 1000-2000 Hz), which in turn shared a significant frequency range with F3 (whose range was about 1800-2800 Hz). When coupled with a 10-15% raising for female voices (due to a correspondingly shorter vocal tract) and even larger changes for children’s voices, the formant overlap ranges across (as well as within) speakers are significant.

In the mid-1970s, the ever-increasing popularity of linear predictive coding (LPC), both for speech coding and speech analysis, displaced formants as the primary speech analysis parameters. Ten years later, the MFCCs took over for ASR, and have remained dominant since then.

We do not propose yet another attempt at formal formant trackers, for two reasons: 1) formant tracking difficulty remains as always, and 2) formants (as such) are not required for ASR. In our opinion, it was an error to insist on a strict formant tracker as a separate module for ASR. Designing independent modules for separate steps in an ASR system may be efficient on a local basis (e.g., allowing different developers), but ignores the interactions and feedback that seem to be prevalent in the speech communication process. Indeed, the HMM approach for ASR succeeded to a certain extent where earlier, step-by-step approaches failed, because of HMM’s global approach, allowing all relevant information to be considered before any serious decision-making. The principle of avoiding an ASR system where each step is taken in isolation is a good idea, but is not a reason to avoid spectral-peak measures such as the formants. Indeed, robust spectral measures better than MFCCs are feasible based on spectral peaks similar to formants, and this is where we propose to raise ASR accuracy. In increasingly noisy speech, the spectral peaks
are the last aspects of speech spectra to be lost. More robust ASR should be possible by directly exploiting peaks instead of approaches that appear to deteriorate quickly in noise.

Our approach is related to the idea of 'missing features' which has appeared recently in the ASR literature [7]. The communication rate ('information' per second) in speech is highly non-uniform, especially in noisy backgrounds, which obscure more the weaker energy at certain times and frequencies than in stronger energy. By putting more weight on 'islands of reliability' in a time-frequency representation, one could hope to raise ASR accuracy [8] more than by the normal HMM procedure that treats all time frames and frequencies equally. (The mel scale, of course, does a simple frequency mapping following some ear behavior, but this is a small part of the nonlinearity of the human speech system.) A low-order spectral representation (more efficient than the MFCCs) is possible that captures the essential aspects of formant structure without requiring formant tracking. Such a measure should focus on spectral peaks, while also discounting overall spectral slope. The latter often varies widely across speakers, speaking conditions, and channel conditions, yet affects phoneme perception little. A peak-based measure can readily exploit this. The MFCCs, on the other hand, are quite affected by spectral slope changes.

The advantage of earlier formant trackers was due to the tendency for formants to merge, split, weaken, or intensify as a complex function of normal speech coarticulation. Trying to reliably track all the formants was a mistaken task for ASR. Identifying the major spectral peaks and their gross dynamics is what appears to be important for ASR; i.e., formants being identified as F1-F3 need not be tracked so rigorously. Instead, we propose a spectral-peak-based analysis measure which can be simultaneously robust, informative, and efficient. Such a measure needs fewer than 10 coefficients to represent the main spectral peaks (their center frequencies and rough bandwidths), thus being more efficient for ASR than MFCCs. Merging and disappearing formants do not cause problems because changes in the slope have strong effects on MFCCs, yet have only weak perceptual effects. A better spectral measure for ASR must be robust to slope changes, and based on relative peak positions, not directly on amplitudes. That is, the presumed importance of the peaks lies not in their amplitudes but in their frequency locations.

6. COMPUTATIONAL CONSIDERATIONS

ASR clearly needs some basic form of spectral analysis (e.g., time-based analysis is too limited), and we assume that a simple FFT can supply the needed information. To simplify comparisons with other ASR methods, we adopt a standard 10-ms-update rate, i.e., do an FFT every 10 ms. For now, we have used a common fixed window length of 25 ms, but we note that this could easily be dynamically varied without negative impact on the algorithms (unlike the HMM method, which requires a fixed update rate). The FFT provides information about the fundamental frequency of the speech (e.g., harmonics are clearly present during voiced speech), which is usually not exploited in ASR of European languages (since F0 is not so critical as it is in tone languages). F0 will nonetheless be helpful for English ASR in the future (subject of future research).

After smoothing the FFT spectrum, simple peak picking is used to locate major peaks (and some approximate bandwidth estimates), without requiring formant labeling. Such processing (smoothing and peak-picking), per frame, is not very costly in computation, compared to the FFT calculation itself.

By way of comparison, consider the computation per frame for basic MFCC-HMM methods for a digit ASR vocabulary. To simplify the situation, assume the following:

a) 10 word-length HMMs (one for each digit-word),
b) 5 states/model (the digits 6 and 7 may well need more states, due to their larger number of phones),
c) 3 Gaussian mixtures/state (three is definitely on the conservative side),
d) 20 parameters/frame (actual systems range widely from 10 MFCCs to more than 40, especially when delta and delta-delta coefficients are included),
e) diagonal covariance matrices (full covariances would significantly increase costs),
f) 20 triangular mel-weighting filters, g) continuous HMMs (vector-quantized (VQ) versions reduce costs, but usually lower ASR accuracy), h) a 512-point FFT (smaller sized FFTs are possible, but only with coarser frequency resolution: 512 points give values every 16 Hz). The number of Gaussian probability density function evaluations per frame can vary widely depending on the choice of algorithm (e.g., forward-backward method, Viterbi method, etc.).

Again simplifying, assume that each parameter vector from a frame of speech is evaluated once in each of the 50 states (5 x 10), and that the Viterbi method is applied, estimating much of the possible calculation that the F-B algorithm uses. Since log-probabilities are often used in calculating joint likelihoods, we will limit the evaluation math to that of the exponential in the Gaussian function (i.e., with the Viterbi approach, summing likelihoods is unnecessary).

Since our method shares the basic FFT with most other approaches (including the standard MFCC-HMM methods), for comparison purposes, we start with the additional calculations beyond the FFT (which includes log conversion to decibels - the perception-based use of log here is well justified). The triangular weighting (to impose mel-scale effects) needs more than 600 M + A operations (as in most DSP cases, each typical operation is a Multiply-Add) - 20 filters with an average of at least 30 taps each. The inverse DFT to produce, say, 10 MFCCs requires about 200 operations (we ignore the ensuing delta-cep operations here).

Our modeling leads to 150 (~3x10^5) PDFs, assuming separate gaussians (unified across models and states). Evaluating each gaussian PDF with a 20-MFCC parameter vector needs 20 subtractions of each parameter mean, 20 squarings, and 20 divisions (normalizing by each variance), followed by 20 additions, to get the desired expo-
7. EXPERIMENTAL RESULTS and DISCUSSION

Using noisy telephone digit strings, our method achieves 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 telephone-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.

8. CONCLUSION

A case can be made that the current HMM-MFCC approach to ASR has sufficient flaws 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.

9. REFERENCES