TEMPORAL SIGNAL PROCESSING FOR ASR

Nelson Morgan

International Computer Science Institute
1947 Center Street
Berkeley, CA 94704

1. INTRODUCTION

For decades, speech recognition systems have used pattern recognition techniques to identify lexical items from a sequence of short-term spectra or cepstra, often with some additional linear or nonlinear processing of these features. This basic scheme owes much to the early methods for channel vocoding; there, filter banks produced short term spectral energies that could be sampled at 50-100 Hz and transmitted to represent the spectral envelope (the excitation information being a separate stream). Both dynamic time warp and HMM approaches to speech recognition could use such a stream of short-term spectral representations to provide a measure of local dissimilarity or similarity to stored examples or models. Much of the effort in front-end signal processing for ASR has been on improving these short-term features.

However, over the last 15 years, approaches have been developed that focus more on the temporal aspect - that is, given some particular feature that varies over time (e.g., 500 Hz signal energy in a 25 ms window), apply signal processing techniques to the sequence of values for that feature. Figure 1 shows the general scheme for such processing. Some of these approaches have become quite standard (for instance, the temporal derivative or "delta" features), while others are new and still quite controversial. In this paper I will review a range of these approaches, closing with what I believe to be some of the most important areas for further work.

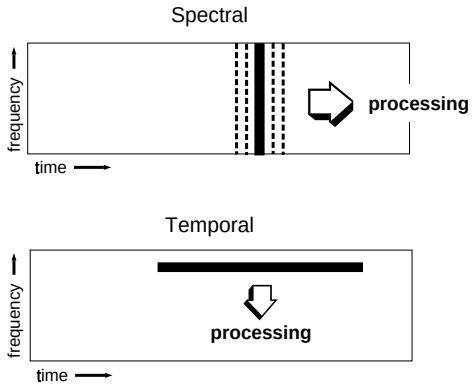

Figure 1: Spectral vs. temporal processing for ASR. In the first case, some measure of the short-term spectrum energy, such as the FFT of a Hamming-windowed signal, is used as the input to processing such as computation of the real cepstrum. This is illustrated in the upper figure. In the second case, the spectral energies (or the processed versions from the first case) are subject to temporal processing. Figure courtesy of Sangita Sharma of OGI.

2. DYNAMIC FEATURES

Feature vectors computed from mel cepstra [7] or PLP analysis [13] correspond to smoothed estimates of local spectra. However, it could be argued that a key characteristic of speech is its dynamic behavior. Because of this, many researchers have made use of estimates of the local time derivatives of the short-term spectrum or cepstrum.

One of the most common forms of this measure is the so-called delta cepstrum [11]. This is typically implemented as a least-squares approximation to the local slope, and as such is a smoother estimate of the local derivative than a simple difference between cepstra for neighboring frames. This can be expressed as

$$\Delta c_i(n) \;=\; \frac{\sum_{k=-N}^{N} k\; c_i(n+k)}{\sum_{k=-N}^{N} k^2} \qquad (1)$$

where a typical value for N is 5.

Thus, each stream of delta cepstral values is computed by correlating the corresponding stream of cepstral values with a straight line that has a slope of 1.

The second derivative (commonly referred to as delta-delta cepstrum) is also often useful, and corresponds to a similar correlation, but with a parabolic function.

Many speech recognition systems have incorporated features such as these. They tend to emphasize the dynamic aspects of the speech spectrum over time, and to be relatively insensitive to constant spectral characteristics that might be unrelated to the linguistic content in speech, such as the long term average spectral slope. However, the resulting feature vectors miss some of the key characteristics that are salient in static spectral representations, and typically are not sufficient for good recognition performance. In practice, most systems that incorporate delta features use them as a complement to static measures such as mel cepstra or PLP cepstra.

Another way of looking at delta or delta-delta features is as a filtered version of the temporal stream for each component of the framewise observation vector. If this choice of temporal processing has often been useful, what about other filtering operations?

3. TEMPORAL FILTERING

The previous section illustrated one of the most common forms of temporal filtering, that of computing local estimates of the time derivative for each component in the sequence of feature vectors. One of the important properties of this measure is that it is insensitive to the average value of each component. Another way to achieve this property would be to calculate a mean vector (a vector whose components are the average of corresponding components in the feature vectors) and subtract it from the feature vectors. If the mean is computed over a sufficiently large chunk of time, the resulting vector sequence will retain the gross characteristics that simple delta filtering can sometimes remove.

In practice, the static features most commonly used for ASR are some form of short term cepstra, so the subtraction of the mean vector is equivalent to a normalization (division) by the geometric mean of the short term power spectra. A little math may clarify this interpretation. Assume that a number of effects (microphone frequency response, effect of turned head of speaker, time-invariant average frequency response of radiation characteristic from speaker's mouth, etc.) can be modeled, to first order, as a linear filtering of the "clean" speech signal. Assume further that a speech signal with short-term spectrum $S(\omega, t)$ is processed by this linear time invariant filter with transfer function $H(\omega, t)$. Then, if $X(\omega, t)$ is the short-term spectrum of the observed signal, we may say

$$X(\omega, t) \;=\; S(\omega, t) H(\omega, t) \qquad (2)$$

Then the corresponding short-term log power spectrum would be

$$log|X(\omega, t)|^2 \;=\; log|S(\omega, t)|^2 \;+\; log|H(\omega, t)|^2 \qquad (3)$$

Thus, a convolutional effect in the time domain (as caused by the filter) corresponds to a multiplication in the frequency domain, and to a sum in the log power domain. If the second factor is relatively constant over the period for which the mean is computed, and if constant components of S are not useful, one can simply estimate the constant component of the sum by computing the mean of the log spectrum. Alternatively, one may compute the Fourier transform of the above components, yielding cepstra, and remove the means in this domain. This operation is a standard one in many speech recognition systems, and is often referred to as Cepstral Mean Subtraction, or CMS (as discussed in [28], along with other approaches to acoustic robustness).

Consider a relevant recognition scenario. A disturbance has affected the speech, and the disturbance might be unknown - a change in telephone channel, a switch in microphones, or perhaps just a turn of the speaker's head so that the overall spectral characteristic is changed. The above analysis suggests that distinguishing between the signal components on the basis of how quickly the cepstrum or log spectrum changes with time can separate out the speech from the convolutional disturbance. In other words, disturbances that were convolutional in the time domain become additive in the log spectral domain. If such additive components have different temporal characteristics, linear filters can be used to separate them out.

Viewed in this more general framework, cepstral mean subtraction can be seen as a specific example of a more general notion of filtering in the domain of the time trajectories of cepstral or log power spectral coefficients. Another specific example of such a principle

is the approach referred to as RASTA-PLP, a modification to PLP analysis that is an on-line approach to achieving robustness to convolutional disturbances [15][18][14].[1] In this approach, the log of each critical band trajectory is filtered with a bandpass filter; typically there is a zero at 0 Hz, and the restriction at the higher frequencies constrains the modulations of log critical band energies to a passband that is required for speech intelligibility. The resulting filtered trajectory is then exponentiated to yield a modified critical band power spectrum for analysis in the later steps of PLP. The use of the log domain for the filtering results in a kind of implicit automatic gain control for the final output sequence.

RASTA filtering can be seen as a generalization of cepstral mean subtraction or delta filtering, both of which apply a linear operation to the temporal sequence for each component of a feature vector. Historically, RASTA processing has either incorporated bandpass filtering between 1 and 12 Hz, or highpass filtering at something like 1 Hz. A related perspective on temporal filtering can be provided by the modulation spectrum [19], which is the spectrum of the energy contour normalized by the average value of the energy envelope. For speech, the higher modulation frequencies (e.g., over 20 Hz or so) have relatively little content; this is the reason why energy envelopes can be sampled at 50-100 Hz for channel vocoders without significant loss in intelligibility. For extremely low frequency bins in the modulation spectrum, the content does not appear to be particularly helpful for speech intelligibility, and in fact is strongly affected by common signal degradations such as the convolutional effects referred to above.

A number of researchers have performed either perceptual tests or ASR experiments to show the relative significance of different parts of the modulation spectrum [8][2][16]. A recently developed signal processing approach is similar to RASTA in spirit, but implements gain control explicitly, separately from the modulation filtering [23].

Finally, temporal filters can be automatically designed using Linear Discriminant Analysis (LDA). LDA finds a linear transformation that maximizes the ratio between between-class variance and within-class variance. In [3], LDA was applied to 1 second of each critical band spectral energy, and the first few eigenvectors of the LDA were used as finite impulse response (FIR) filters to be applied to each time trajectory. The filters essentially do RASTA processing, but their specific characteristics have been designed from data rather than intuition.

[1] The more general idea of filtering temporal trajectories of subband energies, or simple transformations such as cepstral trajectories, is sometimes also called RASTA filtering. RASTA has also been applied to analysis approaches other than PLP; for instance, it has been applied to mel cepstra [26].

4. MULTIPLE-FRAME ANALYSIS

The principal theme of this paper is the processing of temporal sequences of speech features prior to the probability estimation stage. However, the distinction between the two stages can often be blurred. For instance, in the case of hybrid hidden Markov model/artificial neural network (HMM/ANN) systems, the input to the probability estimator (typically a multi-layer perceptron or a recurrent network) consists of multiple sequential feature vectors. During training, the network parameters are adjusted with stochastic gradient descent, and the final weights implement some nonlinear function of the multiple input feature vectors. Thus, both temporal and spectral signal processing are being applied on the path to generating state probabilities. LDA, described earlier, has also been successfully applied to the transformation of variables from one or more frames for speech recognition by a number of researchers in the 1980s and 1990s [20] [12]. A key aspect of these approaches is that features from multiple frames are used to generate probabilities (or likelihoods). Clearly, these time trajectories can be based on speech representations that have been subject to temporal and/or spectral processing. This is the topic of the next two sections.

5. MULTIPLE SPECTRAL STREAMS

In 1993, I participated in the Rutgers Workshop on Conversational Speech Recognition, which was the precursor to the Johns Hopkins workshop of recent years. At that workshop, Hynek Hermansky and I were fortunate to spend a fair amount of time listening to Jont Allen talking about Harvey Fletcher. Allen was a great fan of work that Fletcher had done early in the 20th century. One of the key points that Allen emphasized in these discussions was the temporal perspective - extracting information about phonetic identity from individual spectral channels across time. A key idea (as expounded in [1]), was that unreliable signal components over a limited spectral region would have less of an effect on the ultimate classification than they would for an approach based on full spectral representations computed for each time slice.

Additionally, separate processing of spectral subbands could potentially compensate for variability in the relative timing of phonetic events between different parts of the spectrum. This variability could be exac-

erbated by differing speaking styles (affecting speaking rate, intensity, etc.) and by room reverberation. Such asynchronies were measured and described, for instance, in [24].

Finally, the temporal pattern for some subbands might be more useful for particular phonetic discriminations. For all of these reasons, over the next few years, a number of sites began to experiment with the estimation of probabilities from subsets of the full spectral vector [4][29][6][30][25]. In the early versions of these experiments, improved performance was demonstrated for speech that had been artificially degraded with additive narrowband noise. Following this, improvements were even observed for recognition of clean speech when a subband system was used in combination with a fullband system. Nearly all of these experiments were done with small tasks, although some of them were reasonably realistic. More recent experiments with the large vocabulary Broadcast News task were less conclusive [21].

For these subband experiments, 2-7 bands were used (most commonly 3 or 4). In a recent development, the multi-band recognition approach has been carried to an extreme by using 500-1000 ms of speech over each single critical band energy trajectory. In other words, the temporal filtering strategy described earlier was essentially used to generate a stream of probabilities (or distances) for each filter output channel [17].

It is likely that we still do not know how to properly incorporate subband streams in ASR. Recently, Bourlard has suggested statistical approaches to handle the combination of streams from all possible subsets of a chosen set of subbands, with weights treated as latent variables. We also now know from experiments with Broadcast News that an oracular setting of such weights (setting to 1 for each frame the band that is most likely to give the correct sound as the chosen class) can lead to major improvements in performance [21]. Learning how to set these weights in a more realistic setting may then be the key to wider benefits from multi-band approaches.

6. GENERALIZED MULTI-STREAM ANALYSIS

The previous section described a class of techniques that incorporated a temporal perspective - probabilities were generated separately for different temporal trajectories (of spectral energies). However, the temporal signal processing per se applied to each stream was minimal. In a generalization of this technique, signal processing can be used to generate streams that differ in other aspects than purely spectral. In particular, temporal filtering can be applied to produce multiple spectral representations. Much as the multi-band approach was suggested by the complementarity of different parts of the spectrum for phonetic discrimination, similarly streams with differing temporal properties can potentially complement one another for this purpose.

Probably the oldest (and most successful) application of this approach is one already described in this paper - the use of delta cepstra in addition to the static features. In fact, the delta cepstrum has a different characteristic in the modulation spectrum than does its static equivalent, and consequently has different temporal properties.

More recently, modulation-filtered spectra were combined with PLP at the level of probability streams [21]. This was done in the context of a Broadcast News test. The combination yielded significant improvements in word error rate over the pure PLP case despite the fact that the filtered spectra yielded substantially worse results on their own. In an earlier experiment for a smaller task, we observed a significant complementarity of the errors for each of the two streams in separate decoding passes [31].

This was an ad hoc combination experiment: a "standard" stream was combined with a spectral representation that had a different temporal characteristic, where the latter was developed for some other purpose (to provide robustness under reverberant conditions). What would be a principled criterion for choosing multiple representations, given that they would be combined using a merged statistical model? In one approach, LDA was used as described at the end of Section 3, but was applied to speech for a number of conditions that were not present in the data used for training the statistical models; in particular, several different levels of additive noise and reverberation [27]. These analyses, along with the processing of the "clean data", yielded filters that improved discrimination between target classes (typically phones or syllables). In principle this approach could yield a range of temporal filters that would provide a multi-stream system with representations that would be optimal for a range of potential conditions (in the usual LDA sense of optimality).

Figure 2 shows temporal filters that were generated by this method for the "clean" condition, for phone classes, and for each of 15 critical bands in the telephone bandwidth.

7. DISCUSSION

Researchers have been working to design effective short-term feature vectors for speech recognition for almost

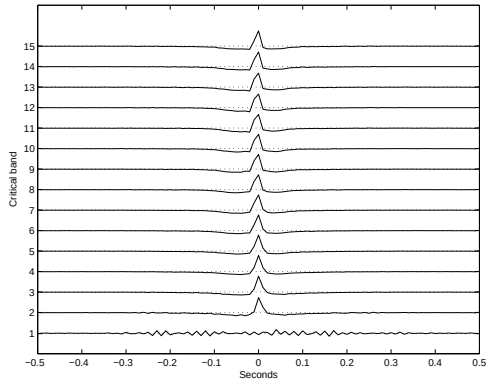

Figure 2: Temporal filters for critical band energies, estimated from a linear discriminant analysis focused on distinguishing between phones in the OGI Stories task [27]. The filter for band 1 is essentially noise, due to the low signal content in this band.

50 years. The history of temporal processing for these features is much shorter. Can powerful statistical methods, both offline and online, make up for the deficiencies in our signal processing methods? Perhaps. However, historically, improvements in the basic signal processing (for instance using cepstral mean subtraction) have provided considerable benefit despite the availability of the statistical mechanisms. In particular, the temporal structure of spectral or cepstral information over time can provide important cues that may not in practice be learned as part of our statistical (acoustic) models. Furthermore, due to the finite size of our training sets and our models, practical systems must always face mismatches between training and test set conditions. Consequently, it may be critical to develop the core signal processing methods much further, if only to provide better raw materials for statistical adaptation and classification during recognition.

Considering temporal processing as well as spectral processing permits a range of possible feature vectors that otherwise would not be considered. It may be that making such a variety of features available to the statistical components will be necessary in order to substantially improve ASR over the current state-of-the-art. Given the computational resources already available, and the likelihood that we will have 100 times as much within a decade, researchers should not be deterred from using many different features in experimental systems. The range of possible variables should include the results from applying different types of temporal processing to spectral or cepstral vectors. Methods such as the LDA-based approach described above should be considered; but also, other criteria for determining temporal or even time-frequency components should be tried. Considered in this light, we have only scratched the surface of plausible measures.

Similarly, we only have the barest understanding of the nature of the speech signal as it is really received for our processing. What is "typical" reverberation for common recognizer use? How does it affect the time-frequency signature of different speech sounds? What is the combined effect of room acoustics, noise, and casual speaking style? What is the interaction between what we have traditionally called language modeling (modeling word sequences), pronunciation modeling (modeling phone sequences), acoustic modeling (modeling state sequences) and front end signal processing? Surely it is an oversimplification that these pieces are modular - for instance, the predictability of words is known to affect their pronunciation [10], and may also have an impact on the utility of different signal processing measures. And as noted in the previous section, it may be the case that some of the advances to come may result from the joint development of acoustic signal processing methods and modifications to the acoustic modeling, two components of ASR that are particularly interdependent.

Thus, there are many questions, and at this point, not so many answers. It is the task of the readers to get to the solutions.

And they won't be found at the back of the book.

Acknowledgements

Over the last 11 years at ICSI, I have been extremely fortunate to work with a fantastic group of both young and senior researchers, many of whom have contributed substantially to the perspective reflected in this paper. My frequent co-authors, Hervé Bourlard and Hynek Hermansky, are of course at the top of this list. Over the last 5 years or so, Steve Greenberg has been another key influence. The students and postdocs have also played a central role in developing the ideas discussed here, with contributions that were particularly relevant to this paper coming from Nikki Mirghafori, Brian Kingsbury, Su-lin Wu, Michael Shire, Dan Ellis, and Takayuki Arai.

Specific thanks are also due to Sangita Sharma of OGI, who helped at the last minute with figures and references so that I could actually finish this paper while traveling. Dr. Sharma was also a strong contributor to many of the technical approaches discussed here. Other editorial assistance was provided by Barry Chen and Elizabeth Weinstein.

And of course, none of this would have been possible without our sponsors; in particular, National Science Foundation grant IRI-9712579 and the European Union SPRACH and RESPITE grants.

8. REFERENCES

[1] Allen, J.B., "How do humans process and recognize speech?," IEEE Trans. on Speech and Audio Processing, vol. 2, no. 4, pp. 567-577, 1994.

[2] Arai, T., and Greenberg, S., "Speech Intelligibility in the Presence of Cross-channel Spectral Asynchrony," Proc. ICASSP '98, Seattle, pp. 933-936, 1999.

[3] Avendano, C., van Vuuren, S., and Hermansky, H., "Data-based RASTA-like filter design for channel normalization in ASR," Proc. ICSLP '96, Philadelphia, vol. 4, pp. 2087-2090, 1996.

[4] Bourlard, H., Dupont, S., Hermansky, H., and Morgan, N., "Towards Subband-based Speech Recognition," Proc. Eusipco '96, Trieste, Italy, 1996.

[5] Bourlard, H., Hermansky, H., and Morgan, N., "Towards Increasing Speech Recognition Error Rates," Speech Communication, vol. 18, pp. 205-231, 1996.

[6] Bourlard, H., and Dupont, S., "A new ASR approach based on independent processing and combination of partial frequency bands," Proc. ICSLP, Philadelphia, pp. 422-425, Oct. 1996.

[7] Davis, S., and Mermelstein, P., "Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357-366, 1980.

[8] Drullman, R., Festen, J.M., and Plomp, R., "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Amer., vol. 95, no. 2, pp. 1053-1064, 1994.

[9] Fletcher, H., Speech and Hearing in Communication, New York: Krieger, 1953.

[10] Fosler-Lussier, E., "Dynamic Pronunciation Models for Automatic Speech Recognition," U.C. Berkeley Ph.D. Dissertation, 1999.

[11] Furui, S., "Speaker independent isolated word recognizer using dynamic features of speech spectrum," IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 34, no. 1, pp. 52-59, 1986.

[12] Haeb-Umbach, R., Geller, D., and Ney, H., "Improvements in connected digit recognition using linear discriminant analysis and mixture densities," Proc. IEEE Intl. Conf. on Acoustics, Speech, & Signal Processing, Adelaide, Australia, pp. II-239-242, 1994.

[13] Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech," Journal Acoust. Soc. Am., vol. 87, no. 4, pp. 1738-1752, 1990.

[14] Hermansky, H., and Morgan, N., "RASTA processing of speech," IEEE Trans. on Speech and Audio Processing, vol. 2, no. 4, pp. 578-589, 1994.

[15] Hermansky, H., Morgan, N., Bayya, A., and Kohn, P., "Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)," Proc. Eurospeech '91, Genova, Italy, pp. 1367-1371, 1991.

[16] Kanadera, N., Arai, T., Hermansky, H., and Pavel, M., "On the importance of various modulation frequencies for speech recognition," Proc. Eurospeech '97, Rhodes, Greece, pp. 1079-1082, 1997.

[17] Hermansky, H., and Sharma, S., "Temporal streams (TRAPS) in ASR noisy speech," Proc. ICASSP, Phoenix, pp. 289-292, Mar. 1999.

[18] Hirsch, H., Meyer, P., and Ruehl, H., "Improved speech recognition using high-pass filtering of subband envelopes," Proc. Eurospeech '91, Genova, Italy, 1991.

[19] Houtgast, T., and Steeneken, H.J.M., "Envelope spectrum and intelligibility of speech in enclosures," Proc. IEEE Conf. on Speech Communication and Processing, pp. 392-395, 1972.

[20] Hunt, M. and Lefebvre, C., "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," Proc. IEEE Intl. Conf. on Acoustics, Speech, & Signal Processing, Glasgow, Scotland, pp. 262-265, 1989.

[21] Janin, A., Ellis, D., and Morgan, N., "Multi-stream speech recognition: Ready for prime time?," Proc. Eurospeech, vol. 2, pp. 591-594, 1999.

[22] Junqua, J.-C., and Haton, J.-P., Robustness in Automatic Speech Recognition, Boston: Kluwer, 1996.

[23] Kingsbury, B.E.D., Morgan, N., and Greenberg, S., "The Modulation-filtered Spectrogram: A Noise-robust Representation," Proc. Robust Methods for Speech Recognition in Adverse Conditions, Tampere, Finland, pp. 95-98, 1999.

[24] Mirghafori, N., and Morgan, N., "Transmissions and transitions: A Study of Two Common Assumptions for Speech Recognition of Natural Numbers," Proc. ICASSP '98, Seattle, pp. 713-716, 1998.

[25] Mirghafori, N., and Morgan, N., "Combining Connectionist Multi-Band and Full-Band Probability Streams for Speech Recognition of Natural Numbers," Proc. ICSLP '98, Sydney, Australia, pp. 743-746, 1998.

[26] Murveit, H., Butzberger, J., and Weintraub, M., "Reduced Channel Dependence for Speech Recognition," Speech and Natural Language Workshop, pp. 280-284, February 23-26, 1992.

[27] Shire, M., "Data-Driven Modulation Filter Design Under Adverse Acoustic Conditions and Using Phonetic and Syllabic Units," Proc. Eurospeech, Budapest, pp. 1123-1126, 1999.

[28] Stern, R., Acero, A., Liu, F.-H., and Oshima, Y., "Signal Processing for Robust Speech Recognition," in Automatic Speech and Speaker recognition, eds. C.H. Lee, F. Soong, and K. Paliwal, Kluwer Academic, 1996.

[29] Tibrewala, S., and Hermansky, H., "Sub-band based recognition of noisy speech," Proc. ICASSP '97, vol. 2, pp. 1255-1258, 1997.

[30] Tomlinson, M.J., Russel, M.J., Moore, R.K., Bucklan, A.P., and Fawley, M.A., "Modelling asynchrony in speech using elementary single-signal decomposition," Proc. ICASSP '97, pp. 1247-1250, April 1997.

[31] Wu, S.-L., Incorporating Information from Syllable-Length Time Scales into Automatic Speech Recognition, U.C. Berkeley Ph.D. Dissertation, 1998.