Joint Distributional Modeling with Cross-Correlation Based Features

Jeff A. Bilmes
<bilmes@cs.berkeley.edu>

International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704, USA

CS Division, Department of EECS
University of California at Berkeley
Berkeley, CA 94720, USA

Abstract - In maximum-likelihood based speech recognition systems, it is important to accurately estimate the joint distribution of feature vectors given a particular acoustic model. In this work, we propose that by modeling the joint distribution of time-localized feature vectors and statistics relating those time-localized feature vectors to the relevant acoustic context, we can estimate information contained in the feature-vector joint distribution without the accompanying theoretical or computational difficulties. We introduce the modcrossgram (MCG), a computational way of estimating short-time spectro-temporal correlation-based statistics that are informative about the feature-vector joint distribution. Using the standard hybrid ANN/HMM architecture, we compare a MCG-based speech recognition system with a more traditional one on an isolated word speech database. We show that, in the presence of noise, the MCG-based system achieves a significant reduction in word error rate over the standard system.

1 Introduction

One important goal in maximum-likelihood based speech recognition systems is obtaining an accurate estimate of the joint distribution of acoustically-derived feature vectors for a particular acoustic model, i.e., $P(X_1^T|M)$, where $\{X_1^T\} = \{X_1, \ldots, X_T\}$, X_t are features representing the t^{th} time frame, and M represents an acoustic model. Fully discriminative modeling [13, 3] notwithstanding, improving the accuracy of the joint distributions $P(X_1^T|M)$ for each model M can often lead to better discriminability between different models.

Hidden Markov models (HMMs) are the most common probabilistic assumptions used for representing the joint distribution. Without any simplifying assumptions,

$$P(X_1^T|M) = \sum_Q P(X_1^T, Q|M) = \sum_Q P(X_1^T|Q, M)P(Q|M),$$

where Q corresponds to the variables comprising the hidden Markov chain. Using the conditional independence assumptions associated with HMMs, the joint distribution can be represented as:

$$P(X_1^T|M) = \sum_Q \prod_t P(X_t|Q_t, M)P(Q_t|Q_{t-1}, M),$$

or under the Viterbi approximation, as:

$$P(X_1^T|M) = \max_Q \frac{1}{c} \prod_t P(X_t|Q_t, M)P(Q_t|Q_{t-1}, M)$$

From IEEE Workshop on Automatic Speech Recognition & Understanding, 1997

where c is a normalizing constant.

One underlying goal in these systems, therefore, is to estimate $P(X_t|Q_t)$,[1] the conditional distribution of the features at time-frame t given the current hidden Markov state variable. Unfortunately, as has been observed [16, 14], such a measure can lead to an inaccurate estimate of the actual distribution of X_t because there is no explicit dependence on the germane acoustic and linguistic context. Statistical regularity contained in natural signals is best represented by the joint distribution of features representing these signals[1] – a model that neglects information about the distribution of X_t that is not contained in Q_t can lead to a distorted representation of this joint distribution.

One approach to this problem is to explicitly model the joint distribution of a short-time window of successive feature frames [13], i.e., $P(X_{t-\ell}^{t+\ell}|Q_t)$. Modeling this joint distribution as such, however, can lead to theoretical difficulties since under most probabilistic assumptions (e.g., HMMs), the joint probability does not factor into a product of probabilities of overlapping feature vectors. Alternatively, one can model the distribution of X_t (or perhaps some segment surrounding X_t) conditioned on additional variables along with the distribution of those additional variables (see [14], and the references contained therein), but this approach adds both theoretical and computational complexity.

As an alternate approach, we propose to model the joint distribution of X_t (which we refer to as *base feature vectors*), and short-time statistics $S(X_t)$, or perhaps some information-preserving reduction thereof, relating X_t to its relevant context – that is, we model $P(X_t, S(X_t)|Q_t)$.

Since an appropriate statistic $S(X_t)$ can contain information about how the current feature vector X_t is related to its surrounding context, the distribution $P(X_t, S(X_t)|Q_t)$ can more accurately model the information contained in the feature-vector joint distribution (i.e., $P(X_{t-d}^{t+d}|Q_t)$ for some d) than can $P(X_t|Q_t)$ alone. On the other hand, rather than modeling $P(X_{t-d}^{t+d}|Q_t)$ explicitly, we model the distribution of parameters, in the form of short-time statistics, representing information about the feature-vector joint distribution.

This approach can have several advantages. First, the parameters representing the statistics $S(X_t)$ may be reduced in a variety of information-preserving ways (e.g., principle component analysis, discrete cosine transform, or something smarter) to obtain a more parsimonious representation of the information contained in the feature-vector joint distribution. Second, the modeling of $S(X_t)$ is simple and is not as bad a violation of an HMM's conditional independence assumptions as is modeling a succession of overlapping feature vectors, since in the former case there is no actual parameter overlap. Third, since $P(X_t, S(X_t)|Q_t) = P(X_t|S(X_t), Q_t)P(S(X_t)|Q_t)$, modeling this joint distribution implicitly conditions on additional variables while simultaneously modeling the distribution of those variables. Thus, a maximum likelihood estimation of this joint distribution is perhaps similar to maximum a posteriori parameter estimation – but rather than determining parameters governing a distribution, we are determining parameters governing the likelihood of statistics which in turn govern a distribution. If the statistics are nearly sufficient, then they are nearly as informative about X_t as is any parameter (such as Q_t) governing X_t's distribution [5] – this, therefore, could reduce the burden normally placed on the variable Q_t to contain all the crucial information about X_t.

While not stated in the same way, others have shown that such an approach can reduce word error in speech recognition systems. For example, one statistic that has shown success in many systems is delta features [8]. In this case, X_t are typically cepstral features, $S(X_t) \approx \frac{d}{dt}X_t$, and we model an approximation of the joint distribution of base features and their temporal derivatives, or $P(X_t, \frac{d}{dt}X_t|Q_t)$.

In looking at the relation between two feature vectors, one simple but informative statistic is the correlation matrix $R_{xy} = E[XY^T]$. This second order statistic can describe, at some

[1] We henceforth omit the conditional dependence on the model M.

level, information about the joint distribution of X and Y. Therefore, we propose to augment our probabilistic model with new features representing the statistical correlation between base features. Furthermore, we observe the correlation both across time and across feature position to capture temporal, positional (typically frequency or quefrency), and temporo-positional dependencies. Our model therefore consists of the following:

$$P(X_t, \frac{d}{dt}X_t, \bigcup_{l \in C_t} E_s[X_t X_l^T] | Q_t) \tag{1}$$

where C_t specifies the relevant context window around t, and $E_s[]$ is an expected value estimate over a short time duration of length s.

Speech, like other naturally occurring signals (including noise), has unique statistics[2]. Directly modeling the distribution of speech statistics, such as correlations, might therefore lead to a system less sensitive to noise whose statistics are not speech-like. White noise is an obvious example, but it is likely that more complex noise with non-speech-like statistics will have a weaker effect on a system specifically tuned to speech-like statistics. For example, a particular set of correlation pairs at a given time point might strongly indicate a certain speech sound. It is unlikely, however, that noise with non-speech-like statistics will have exactly that set of correlations at that time point. While any system that models the feature-vector joint distribution will have access to the information provided by such correlational cues, in our model we avoid the complexity and retain the correlational information of those systems. Furthermore, it may be possible to model the statistics of individual speech sounds (such as phones, syllables, etc.) by observing only those locally pertinent correlational features.

The use of such correlational features might resemble actual processing performed by perceptual systems. Co-modulation masking release (CMR) is a perceptual phenomenon where statistical coherence across critical-band channels can cause a significant decrease of the detectability threshold of a narrow-band signal contained in one of the channels. It is believed, therefore, that some form of cross-channel correlation of modulation-envelope-like signals occurs in the mammalian auditory pathways [15, 10]. Furthermore, auditory echoic memory [12] is a phenomenon whereby a buffer of approximately 200ms apparently exists in the auditory system holding pre-perceptual information before subsequent higher-level processing takes place. Echoic memory indicates a storage arena from which cross-channel correlations could be "computed", and the CMR effect indicates one possible use of such correlations. Another possible use is as follows: It is believed that perceptual systems have adapted to efficiently encode information about natural scenes in their environment [1, 7]. An efficient encoding requires information about a joint distribution – signals believed likely according to this distribution are encoded more efficiently (i.e., with fewer bits). The use of correlational features could therefore be used, in lieu of the actual joint distribution, as a cheap way to make decisions about the encoding.

In Section 2, we introduce the *modcrossgram*, a way of computationally representing short-time correlational features. In Section 3, we demonstrate that using such features in a speech recognition system can result in a significant word error decrease in a noisy test case. In Section 4, we attempt to visually depict our representation. And in Section 5, we conclude and give current and future work.

2 The ModCrossGram: An Acoustic Approach

The *modcrossgram* (*mod*ulation envelopes *cross*-correlated, or MCG) explicitly computes the short-time cross-correlation between disparate regions of the spectrum of modulation envelopes. The resulting features can then be directly incorporated into any standard speech recognition system.

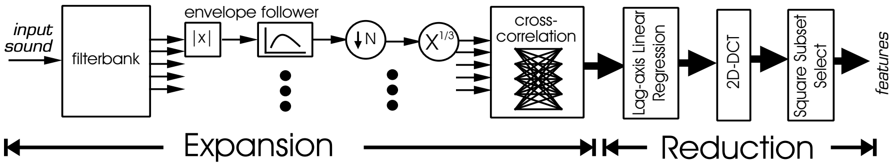

Figure 1: The expansion and reduction stages of the MCG procedure.

The MCG computation consists of two stages: the first is an expansion that computes all cross-correlations within a given range, and the second is a reduction that decreases the number of parameters while attempting to minimize information loss.

The expansion stage is as follows (see Figure 1): we first compute the modulation envelopes in each channel of a critical-band-like filter-bank by rectifying and band-pass filtering the filter-bank outputs. A band-pass filter is used to remove low- and high-frequency modulation energy known to be of little importance to speech intelligibility [6, 11]. The dynamic range is subsequently reduced by cube-root processing.

The modulation envelopes are then processed by short-time cross-correlation defined as:

$$R_{i,j}(t,\ell) = \sum_{k=0}^{N} x_i(t+k)x_j(t+k+\ell)w_k,$$

where x_i is the i^{th} envelope channel, t is the starting offset within the signals, ℓ is the correlation lag, $N+1$ is the number of points used to compute the correlation which we call the *correlation window* (corresponding to s in $E_s[]$ of Equation 1), and w_k are windowing coefficients. ℓ ranges over a time-span we call the *context window* and corresponds to the range of C_t in Equation 1. All pairs of channels are processed at each time step. The result is a four-dimensional representation of the original signal – for each t, a rectangular prism whose three principle axes are indexed respectively by the two frequency channels, and the correlation lag.

The reduction stage is as follows: first, a linear-regression is applied in the lag-dimension, representing the lag axis of each frequency channel pair as a single numerical slope. While retaining the temporal trend of the correlations, this step removes DC information that, as in cepstral analysis, is found to be of less value. A two-dimensional DCT is then applied to the resulting matrix and a low-frequency square subset is taken as the ultimate feature set.

The MCG, therefore, provides a means to compute features representing the short-time cross-correlation between different spectro-temporal regions in a source signal. Such features might, therefore, help HMM-based recognition systems better estimate crucial information about the joint distribution.

3 Speech recognition results

We evaluated MCG based features in a standard hybrid artificial neural network/hidden Markov model (ANN/HMM) speech recognition system [13] using a telephone quality database of isolated digits and control words from Bellcore. We compared these results with those from a more traditional system. Each system was tested using data from 200 speakers totaling 2600 examples from 4 jackknifed cuts – scores shown are the average of 4 tests in which 150 speakers were used for training and 50 different speakers used for testing. We tested both using clean data and data with 10db SNR additive automobile noise recorded over a cellular phone. The training procedure included only the clean data. Before ANN training, all weights were set to small random values.

We used jrasta [11] for the base features and their derivatives, but the MCG, as shown in Figure 1, was derived from compressed envelopes. Therefore, the actual probabilistic model is as shown in Equation 2 where $R(\cdot)$ is the jrasta transformation of envelope based features, and $M(\cdot)$ is the MCG reduction as described in the previous section.

$$P(R(X_t), \frac{d}{dt}R(X_t), M(\bigcup_{l \in \mathcal{C}_t} E_s[X_t X_l^T])|Q_t) \tag{2}$$

Each frame of base features and derivatives consists of 8 jrasta values plus 9 deltas (log energy is not used). The standard jrasta best case uses 9 frames of context as input to an ANN – previous experiments have shown no benefit to using longer or shorter jrasta contexts.

The MCG filter-bank produced 22 quarter-octave channels using FIR filters designed by a Kaiser windowing method. The envelope band-pass filter, again Kaiser-FIR, restricted the modulation envelope energy to a range between approximately 1 and 35Hz. The down-sampling factor was 100 giving the sub-band envelopes a sampling rate of 80Hz. The cross-correlation stage used a 50ms (four 12.5ms frames) correlation window, a 212.5ms (17 frames) context window, and a rectangular shaped set of windowing coefficients. The final features consisted of an 11x11 subset of the 2D-DCT output.

We compare three systems in Table 1. The first uses 9 jrasta frames of context only, the second uses 1 jrasta frame of context only, and the third combines a single jrasta frame with a frame of 121 MCG-based features.

Features used	NIU/NHU	Clean	10dbSNR
9 Jrasta Frames	153/200	1.63	10.73
1 Jrasta Frame	17/572	3.70	15.48
MCG + 1 Jrasta Frame	138/216	1.88	8.35

Table 1: Word error results for the three systems. The columns, from left to right, show the features used in that system, the number of ANN input units (NIU) and hidden units (NHU), the clean-case word error rate (WER), and the WER in the presence of noise. In each system, there are 56 output units corresponding to the number of possible phones. The number of hidden units is set in each experiment to equalize the number of free parameters.

As Table 1 shows, the use of nine rather than one frame of jrasta features results in a significant decrease in word error rate both in the clean and the noisy test case[2]. The combined MCG/jrasta system results in an insignificant word-error difference in the clean test case with the 9-frame jrasta system. The MCG system, however, shows a significant word error decrease over jrasta alone ($p < 0.002$ using a difference of proportions significance test) in the noisy test condition. This value, 8.35%, is insignificantly different then the overall best error rate our group has achieved for this condition [4] which requires more training parameters and a more complicated recognition process.

The jrasta features were designed for noise robustness but the MCG recognizer wins nonetheless, probably because, as discussed earlier, modeling the joint distribution between X_t and the correlation based features results in a system that is more statistically blind to such noise.

While we tested a variety of lengths for the correlation and context window, it is interesting to note that the resulting "best" values, 50ms and 212.5ms respectively, are close to the 40ms time quanta and 200ms integration period that may be crucial time-constants in the mammalian auditory system [9].

[2] These jrasta-only results are slightly better than those reported in previous papers as the experiments were re-run and re-tuned to a new recognition system.

4 Visual Example

In this section, through visual display, we offer an explanation as to how MCG based features could help recognition in the presence of noise.

Figure 2 shows 3 plots. The top plot shows the spectrogram of the utterance “ka ga”, two possibly confusible syllables. As the spectrogram shows, these two syllables differ principally in the voicing onset time – the delay between the stop-burst and the start of the periodic voicing – which is about 80ms in “ka” but nearly simultaneous in “ga”.

The middle plot shows the fully expanded 4-dimensional MCG for the utterance projected down onto two dimensions and time-aligned to the spectrogram. This picture shows the top 60dB of the positive correlations computed using a context window of 212.5ms and correlation window of 50ms. At each frame number and frequency channel, a small matrix shows the correlation between that channel and all other channels (vertical axis) and the lag from -8 frames to 8 frames (horizontal axis). While it is perhaps interesting to view the full MCG, more intuition can be gained by viewing a only subset of the features.

The bottom plot shows the top 60db of the positive MCG between channel 17 (center frequency 1560Hz) and 8 (center frequency 328Hz). Each row shows the correlation between these two frequency channels at the corresponding lag and time frame. Observe that for “ka”, at around frame 12, we see significant correlation at positive lags of 70-80ms; a little later (frame 19), we see significant correlation at the corresponding negative lags. This reflects the timing difference between the initial stop release and the subsequent voiced onset. As expected, these correlations are not observed for “ga”, which exhibits quite different patterns around the corresponding frames, 65 and 70.

This simple example shows that MCG features can capture certain second order statistical properties of natural speech categories (in this case, the correlation at the beginning of the syllable “ka” between certain frequency channels at certain time lags). This information is unlikely to exist in a single frame of base feature vectors. Furthermore, it is unlikely that noise will have a strong correlation at this position. In other words, an indication of this particular speech sound is contained in its pattern of redundancy and it is unlikely that noise or some other speech sound will contain this exact pattern of redundancy. The correlational based approach of the MCG explicitly models this redundancy and can therefore, at an acoustic level, utilize cues that normally are available only via the feature-vector joint distribution.

5 Conclusions, Current and Future Work

We have shown that modeling the joint distribution between feature vectors and statistics relating those feature vectors to their relevant context can benefit isolated-word speech recognition. In particular, the additional use of correlation-base features, along with base features and their deltas, results in a significant decrease in word error rate in a 10dB SNR noisy test case.

We are currently investigating the use of statistical regularity derived directly from the training data to determine better MCG reduction strategies. One approach based on maximum mutual information seems promising. We also plan to test other reduction strategies – one method, potentially useful for speaker normalization, reduces across correlation pairs with a constant log-frequency difference. We also plan to evaluate our approach using a fully likelihood-based system such as HTK. Since a primary goal of this work is to more accurately model the information contained in the joint distribution of features of a speech signal, we also plan to investigate other statistical methods to this end.

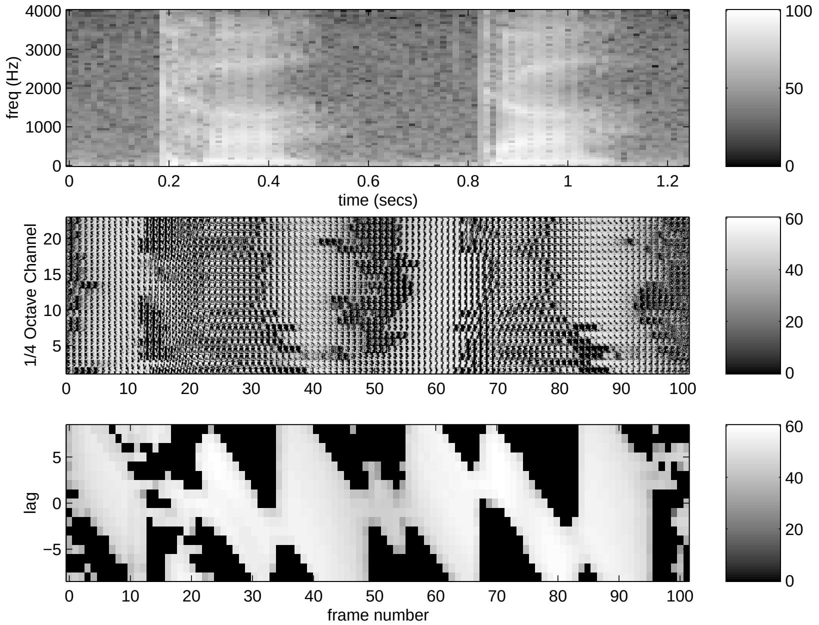

Figure 2: Top: Spectrogram of the utterance “ka ga”. Middle: fully expanded MCG. Bottom: MCG subset indicating cross-correlation between channels 17 (CF 1560 Hz) and 8 (CF 328 Hz).

6 Acknowledgments

This work has benefited from many useful discussions with Nelson Morgan, Dan Ellis, Steve Greenberg, and other members of the speech group at ICSI. This work has been partially sponsored by JSEP contract F49620-94-C-0038 and ONR URI Grant N00014-92-J-1617.

References

[1] J.J. Atick. Could information theory provide an ecological theory of sensory processing. *Network*, 3:213–251, 1992.

[2] H. Attias and C.E. Schreiner. Temporal low-order statistics of natural sounds. *NIPS*, 9, 1997.

[3] L.R. Bahl, P.F. Brown, P.V. de Souza, and R.L. Mercer. Maximum mutual information estimation of hidden markov model parameters for speech recognition. In *Proc. ICASSP*, pages 49–52, Tokyo, 1986.

[4] J. Bilmes, N. Morgan, S.-L. Wu, and H. Bourlard. Stochastic perceptual speech models with durational dependence. *ICSLP*, November 1996.

[5] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[6] R. Drullman, J.M. Festen, and R. Plomp. Effect of reducing slow temporal modulations on speech reception. *JASA*, 95(5):2670–2680, May 1994.

[7] R. Fieke, D.A. Bodnar, and W. Bialek. Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proc. R. Soc. Lond. B*, 262:259–265, 1995.

[8] S. Furui. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Trans. ASSP*, 34(1):52–59, February 1986.

[9] S. Greenberg, D. Poeppel, and T. Roberts. A space-time theory of pitch and timbre based on cortical expansion of the cochlear traveling wave delay. In *Proc. 11th Int. Symp. on Hearing*, pages 257–262, Grantham, U.K., August 1997.

[10] J.W. Hall, J.H. Grose, and L. Mendoza. Across-channel processes in masking. In B.C.J. Moore, editor, *Hearing*, Handbook of Perception and Cognition, chapter 7, pages 243–266. Adacemic Press, 2nd ed. edition, 1995.

[11] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Trans. Speech and Audio Proc.*, 2(4):578–589, October 1994.

[12] D.W. Massaro. Preperceptual auditory images. *J. of Exp. Psych.*, 85(3):411–417, 1970.

[13] N. Morgan and H. Bourlard. Continuous speech recognition. *IEEE Signal Processing Magazine*, 12(3), May 1995.

[14] M. Ostendorf, V. Digalakis, and O. Kimball. From HMM's to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Trans. Speech and Audio Proc.*, 4(5):360–378, September 1996.

[15] V.M. Richards. Monaural envelope correlation perception. *JASA*, 82(5):1621–1630, November 1987.

[16] C.J. Wellekens. Explicit time correlation in hidden markov models for speech recognition. *ICASSP*, pages 384–386, 1987.