A NOVEL APPROACH USING MODULATION FEATURES FOR MULTIPHONE-BASED SPEECH RECOGNITION

Pascal Clark, Gregory Sell**, and Les Atlas**

*Department of Electrical Engineering, University of Washington, Seattle, WA
**Center for Computer Research in Music and Acoustics, Stanford University
{clarkcp, atlas}@u.washington.edu, gsell@ccrma.stanford.edu

ABSTRACT

Recent advances in coherent and convex demodulation have proven useful for analyzing and modifying the low-frequency envelope structure of speech. This paper reports the application of both methods, referred to here as bandwidth-constrained demodulation, to large-scale speech recognition in the form of new feature representations. Modulation-based features yielded measurable improvement when included as complementary sources of information with a baseline recognizer. Furthermore, both sets of demodulation features showed promise for outperforming the conventional Hilbert envelope method which underlies most modern speech recognition features. These experimental results show the potential for further development in feature representations based on recently-developed bandwidth-constrained modulation signal models.

Index Terms— Modulation, Convex optimization, Feature extraction, Speech recognition, Speech processing

1. INTRODUCTION

This paper summarizes the results of new modulation-based acoustic features for automatic speech recognition (ASR), completed during the 2010 summer workshop hosted by the Center for Language and Speech Processing at Johns Hopkins University. The focus of the workshop was segmental conditional random fields (SCRF), which are notable for their ability to integrate multiple classifiers at the word level [1]. Within this framework we combined our experimental features with a well-established hidden Markov model (HMM) word detector. Thus we had the opportunity to complement, rather than compete with, an existing ASR system while testing theoretical predictions related to foundational concepts of modulation representations for speech.

Modulation-based features, in the form of the spectrogram and mel-frequency cepstral coefficients (MFCC), have underpinned speech recognition since as early as the mid twentieth century. Short-time Fourier coefficients are equivalent to subband-amplitude signals, the magnitudes of which correspond to a method of demodulation called the Hilbert envelope. Modern ASR features have since added linear, nonlinear, discriminative and speaker-adaptive transforms for improved classification, but fundamentally begin with Hilbert envelopes.

Generalizing demodulation in terms of a signal-product model, however, reveals that the Hilbert envelope is an arbitrary solution to an under-determined problem. Different constraints on the model can therefore lead to better-behaved results, as developed in the form of coherent [2] and convex [3] demodulation. Alternative methods of demodulation raise the possibility of building a firmer foundation, other than the Hilbert envelope, for future development of informative ASR features. With that in mind, the point of this work is to demonstrate the viability of bandwidth-constrained demodulation features in a large-scale speech recognition system.

Other methods of modulation-based speech recognition have focused on modifying the Hilbert envelope. Notable examples are modulation filtering [4][5][6] and frequency-domain linear prediction [7]. We instead estimate modulator signals as solutions to a constrained product-model synthesis equation. In convex demodulation this takes the form of an optimization problem, while coherent demodulation is based on signal-adaptive carrier estimation. In this paper we present new extensions on convex and coherent demodulation algorithms for the purpose of speech recognition.

We begin by framing the speech classification problem in Section 2 and then define the speech modulation signal model in Section 3. We describe two novel demodulation methods, compared to the Hilbert envelope, in Section 4 as the first step toward the template-based multiphone classification system outlined in Section 5. Finally, we discuss experimental results in Section 6 and conclude in Section 7.

2. CLASSIFICATION BACKGROUND

In automatic speech recognition the task is to identify a linguistic utterance - a word, syllable, or phoneme - using

We acknowledge support of the JHU Summer Workshop and AFOSR Grant FA9550-09-1-0060.

slowly-varying local features of the acoustic data. Let $x[n]$ be a time-domain speech signal sampled at rate f_s. For K features we compute the vector expansion $M[k,i]$ in the neighborhood around $n = Ri$,

$$M[k,i] = F\{h[Ri-n]x[n]\} \quad (1)$$

where $h[n]$ is a finite window function and R is an integer downsampling factor. Given a $K \times I$ matrix of concatenated feature vectors $\boldsymbol{M}$, a direct classifier chooses a label w according to the maximum a posteriori criterion

$$\hat{w} = \underset{w \in W}{\operatorname{argmax}}\, p(w|\boldsymbol{M}) \quad (2)$$

where W is a lexicon of possible utterances and the probability model $p(w|\boldsymbol{M})$ is parametrically fitted to training data.

The success of the classifier in (2) depends principally on two things: a) correct characterization of the probability models for all $w \in W$, and b) the design of $F\{\cdot\}$ such that the K-dimensional distributions $p(\boldsymbol{M}|w_i)p(w_i)$ and $p(\boldsymbol{M}|w_j)p(w_j)$ are disjoint in feature-space for all $i \neq j$. Clearly both a) and b) are important, but in this paper our focus is the design of informative features $M[k,i]$ using the principles of bandwidth-constrained demodulation.

If $F\{\cdot\}$ is the discrete Fourier transform followed by the magnitude operation, then (1) yields the conventional Hilbert envelope representation. In the next section, we generalize demodulation in terms of a signal-product model for speech.

3. SPEECH MODULATION SIGNAL MODEL

The first main contribution of this paper is to propose a framework for estimating features $M[k,i]$ based on the vector expansion $m_k[n]$ satisfying the sum-of-products model [2][3]:

$$x[n] = \sum_{k=0}^{K-1} m_k[n] \cdot c_k[n] \quad (3)$$

where the dot indicates sample-wise multiplication and K is a finite integer. In this model, the modulators $m_k[n]$ each vary slowly with n while the quickly-oscillating carriers $c_k[n]$ serve primarily to frequency-shift baseband modulations into the acoustic range of hearing. We assume that the modulators contain necessary cues for understanding speech, at frequencies around the syllabic and phonetic rates.

Further assuming that the signal products $s_k[n] = m_k[n] \cdot c_k[n]$ are bandpass and spectrally non-overlapping, we define $s_k[n]$ as the output of a bandpass, possibly time-varying, filter operation [8]

$$s_k[n] = \sum_{\tau} x[\tau] h_k[n, n-\tau]. \quad (4)$$

The problem of estimating $m_k[n]$ from $s_k[n]$ is called *demodulation* and is treated in more detail in the next section. For now we emphasize that there is no unique solution for $m_k[n]$ without further constraints, for the same reason that any number a has no unique factorization (b, c) such that $a = b \cdot c$.

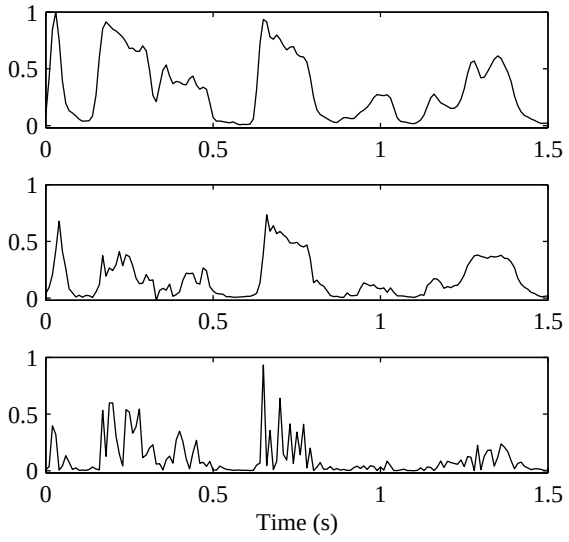

Fig. 1. Example modulator waveforms corresponding to the 0-500 Hz subband from male speech "The thing about bird populations"). From top to bottom: convex, coherent, and Hilbert envelopes.

Convex and coherent demodulation each make explicit assumptions on (3) as a means toward emphasizing low-frequency modulations without harmonic pitch interference. One type of constraint is the definition of $c_k[n]$, whose inclusion in (3) may appear superfluous except for the fact that the characteristics of $c_k[n]$ exactly complement those of $m_k[n]$. In bandwidth-constrained demodulation, carrier constraints absorb non-syllabic fine structure, such as pitch oscillations, so as to leave linguistic cues undisturbed in the modulators. This is equivalent to enforcing smoothness across frame time-locations i in the feature array $M[k,i]$, as discussed next.

4. DEMODULATION METHODS

To reduce computational load during training and classification, we require feature vectors that are low-dimensional and decimated in time. This becomes somewhat of a problem with respect to subband demodulation, because bandwidth is inversely proportional to K and broader bands can contain multiple interfering harmonics. In the following we describe how bandwidth-constrained demodulation mitigates such interference while maintaining a low-dimensional (small K) representation, compared to the non-mitigated Hilbert envelope. A visual comparison also appears in Figure 1.

4.1. Convex Demodulation

Here we pose the demodulation task as an optimization problem [3]. Defining $h_k[n,\tau] = h_k[\tau]$ to be a time-invariant filter, the optimal modulator for a given subband signal $s_k[n]$ is one which minimizes high modulation frequencies subject to signal-dependent amplitude constraints.

In this paper we present a new frequency-domain version equivalent to, but faster to solve than, the linear method in [3]. Specifically, we find the real-valued modulator coefficients θ_l which solve the following convex problem:

$$\begin{array}{ll} \text{minimize} & \boldsymbol{\theta}^T \left(\boldsymbol{W}\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{W} + \boldsymbol{B}^T\boldsymbol{B} \right) \boldsymbol{\theta} \\ \text{subject to} & m_k[n] = \sum_l \theta_l b_l[n] \\ & |s_k[n]| - m_k[n] \le 0, \quad n \in P \end{array} \tag{5}$$

where $\boldsymbol{B}$ is a basis matrix of cosine and sine functions $b_l[n]$, $\boldsymbol{W}$ is a highpass diagonal matrix, and P is the set of indices for which $|s_k[n]|$ has a local maximum. The implicit carrier constraint here is $c_k[n] = 1$ for $n \in P$, since (5) smoothly interpolates the kth modulator between the local maxima of $|s_k[n]|$.

The resulting feature-vector time series is then

$$M_{CVX}[k,i] = m_k[Ri]. \tag{6}$$

4.2. Pitch-Invariant Coherent Demodulation

Unlike its convex counterpart, coherent demodulation defines adaptive subband signals centered on finite-bandwidth time-varying sinusoids [2]. We assume harmonic carriers:

$$\begin{aligned} c_k[n] &= \exp(jk\phi_0[n]), \quad 0 \le k < K' \\ m_k[n] &= \sum_\tau \left(x[\tau] \cdot c_k^*[\tau] \right) h[n-\tau] \end{aligned} \tag{7}$$

where superscript * denotes complex conjugation, $\phi_0[n]$ is radian phase corresponding to the fundamental frequency $F_0[n]$, and $h[n]$ is a time-invariant lowpass filter that limits the modulator bandwidth. Assuming $F_0[n]$ varies slowly, the second line of (7) approximates a basebanded version of (4). See [9] for details.

To eliminate pitch-dependent variation in $m_k[n]$, we introduce a new, pitch-invariant extension to [2]. Specifically, we treat $m_k[n]$ as K' samples of an underlying transfer function at time n, and resample the k-axis to a constant reference "pitch" of $F_{ref} = f_s/2K$. For this application we choose a large K' so that the carriers cover the spectrum, and then resample by a factor of $F_0[n]/F_{ref}$.

With the vector resampling operator $T\{\vec{\boldsymbol{m}}, F_0, F_{ref}\}$, the feature-vector time series is

$$M_{COH}[k,i] = T\{|m_k[Ri]|, F_0[Ri], f_s/2K\}. \tag{8}$$

Although $m_k[n]$ is complex-valued, we use only the magnitudes because of the absence of consistent structure in the modulator phase.

4.3. Hilbert Envelope Demodulation

To complete our comparison in the upcoming speech recognition experiments, we also include the conventional Hilbert envelope method. Hilbert modulators and carriers are typically defined with respect to fixed subband signals such that

$$\begin{aligned} c_k[n] &= \exp\{j \arg(s_k[n])\}, \quad 0 \le k < K \\ m_k[n] &= |s_k[n]| \end{aligned} \tag{9}$$

where $s_k[n]$ is an analytic subband from a complex, time-invariant filter $h_k[\tau]$. The corresponding feature vectors are then

$$M_{HIL}[k,i] = m_k[Ri]. \tag{10}$$

Unlike convex demodulation the modulators are not smoothed, and unlike coherent demodulation the subbands are not signal adaptive. For broadband $s_k[n]$ this means that the resulting $m_k[n]$ will contain harmonic cross-terms in the form of high-frequency modulations, which alias after downsampling by R in a signal-dependent way.

5. MULTIPHONE DISCRIMINATION WITH MODULATION TEMPLATES

To take advantage of the temporal bandwidth constraints on our demodulation features, we defined a classification lexicon of multi-phonetic sequences using the maximum mutual information (MMI) technique in [10]. We avoided segmentation issues by restricting our lexicon W to the 607 MMI multiphones which are also full words. For each multiphone w_i we trained a discriminative template $\vec{\boldsymbol{\Lambda}}_{\boldsymbol{i}}$ to evaluate the likelihood ratio $\ell_i(\boldsymbol{M}) = p(w_i|\boldsymbol{M})/p(w_i^c|\boldsymbol{M})$, where w_i^c denotes the set of all multiphones except w_i. Let $\vec{\boldsymbol{M}} = M[iK+k]$ be the vector version of a feature matrix $\boldsymbol{M}$. Using maximum-entropy models [11] we represented likelihoods of the form

$$\ell_i(\boldsymbol{M}) = \exp\left(\vec{\boldsymbol{\Lambda}}_i^T \left[\vec{\boldsymbol{M}}^2; \vec{\boldsymbol{M}}; 1 \right] \right) \tag{11}$$

where $\vec{\boldsymbol{\Lambda}}_{\boldsymbol{i}}$ is a vector template of length $2KI+1$, $\vec{\boldsymbol{M}}^2$ is element-wise squared, and $[;]$ denotes vertical concatenation.

Figure 2 plots multiphone classification error rates obtained from the Broadcast News corpus with this method. Lines of regression demonstrate that, on average, convex features perform better than Hilbert features (with a slope of 0.88). The error-rate spread for coherent templates, on the other hand, generally shows poorer classification performance compared to Hilbert templates. Although informative, these comparisons do not necessarily relate to how the features will perform on a multi-word segment level as modeled by an SCRF, which is what we explore in the next section.

6. SETUP AND RESULTS

Treating $\ell_i(\boldsymbol{M})$ as the score for the word hypothesis w_i, we annotated a baseline lattice generated by the IBM Attila

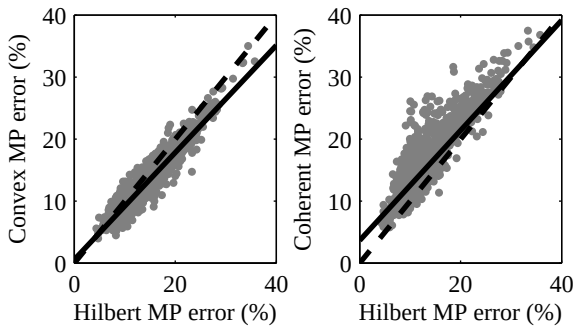

Fig. 2. Multiphone classification error rates overlaid with lines of regression (solid) and lines of equal error (dashed). Note that chance is 50% since each multiphone classifier makes a yes/no decision.

decoder [12] and fed the result into an SCRF-based speech recognizer implemented by the SCARF toolkit. We trained on about 430 hours of 16 kHz audio from the Broadcast News corpus, and obtained recognition scores by decoding the NIST dev04 set of about 22k words. See [13] for details on the full setup.

For both convex and Hilbert envelope demodulation, we chose a uniform subband width of 500 Hz, which [14] determined to be the maximum bandwidth without sacrificing speech information in the modulators. Likewise, we resampled 30 harmonics in the coherent method to a reference pitch of 500 Hz, so that all demodulation methods resulted in 16-dimensional feature vectors. The modulation frequency cutoffs were 30 Hz for convex demodulation and 50 Hz for coherent. The time-decimation factor R was 160 which resulted in a modulation sampling rate of 100 Hz.

We trained SCRF models using four annotation methods: 1) non-annotated, 2) convex-modulation scores, 3) coherent-modulation scores and 4) Hilbert envelope scores. In each case, we incorporated the one-best HMM sequence from Attila as a baseline feature and used a trigram language model. The resulting word-error rates (WER) changed by about -0.2% for both convex and coherent annotations relative to the non-annotated WER of 16.0%. Hilbert annotations, on the other hand, resulted in a smaller change of -0.1%.

7. CONCLUSION

In a large scale speech recognition task, these early results demonstrate the viability of recently-developed modulation-based features for multiphone recognition. The modulation features complemented a state-of-the-art baseline system within an SCRF framework in order to reduce word-error rate by an absolute 0.2%. Furthermore, the results indicate that bandwidth-constrained demodulation can perform better than the conventional Hilbert envelope which underlies most modern ASR features. Our bandwidth-constrained modulators offer a starting point for further development in dimensional reduction, discriminative transforms and speaker adaptation. These results thus open the possibility for new representations of low frequency envelope information in speech recognition systems.

We are grateful to Geoffrey Zweig, Patrick Nguyen and Damianos Karakos for their invaluable expertise, and we dedicate this work to the late Prof. Fred Jelinek.

8. REFERENCES

[1] G. Zweig and P. Nguyen, "A segmental CRF approach to large vocabulary continuous speech recognition," in *Proc. ASRU*, 2009.

[2] P. Clark and L. Atlas, "Time-frequency coherent modulation filtering of nonstationary signals," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4323–4332, Nov. 2009.

[3] G. Sell and M. Slaney, "Solving demodulation as an optimization problem," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2051 –2066, Nov. 2010.

[4] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech, Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[5] B.E.D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, no. 1-3, pp. 117 – 132, 1998.

[6] Y.-H.B. Chiu and R.M. Stern, "Minimum variance modulation filter for robust speech recognition," in *Proc. IEEE ICASSP, Taipei*, April 2009, pp. 3917 –3920.

[7] S. Thomas, S. Ganapathy, and H. Hermansky, "Phoneme recognition using spectral envelope and modulation frequency features," in *Proc. IEEE ICASSP*, April 2009, pp. 4453–4456.

[8] T. Kailath, *Channel characterization: Time-variant dispersive channels*, McGraw-Hill, New York, NY, 1961.

[9] L. Atlas, P. Clark, and S. Schimmel, "Modulation Toolbox version 2.1 for MATLAB," http://isdl.ee.washington.edu/projects/modulationtoolbox/, Sept. 2010.

[10] G. Zweig and P. Nguyen, "Maximum mutual information multi-phone units in direct modeling," in *Proc. Interspeech*, 2009.

[11] S.F. Chen and R. Rosenfeld, "A survey of smoothing techniques for me models," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 1, pp. 37 –50, Jan. 2000.

[12] S.F. Chen *et al.*, "Advances in speech transcription at IBM under the DARPA EARS program," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1596 –1608, Sept. 2006.

[13] G. Zweig *et al.*, "Speech recognition with segmental conditional random fields: A summary of the JHU 2010 Summer Workshop," Submitted to IEEE ICASSP 2011.

[14] G. Sell and M. Slaney, "The information content of demodulated speech," in *Proc. IEEE ICASSP*, Dallas, TX, March 2010, pp. 5470–5473.