THE INFORMATION CONTENT OF DEMODULATED SPEECH

Gregory Sell[1] *

Malcolm Slaney[2,1]

[1]Center for Computer Research in Music and Acoustics
Stanford University
gsell@ccrma.stanford.edu

[2]Yahoo! Research
Sunnyvale, CA 94089
malcolm@ieee.org

ABSTRACT

In this paper we describe the effect of demodulation on speech signals. We compare two different algorithms for demodulating audio: the classic approach based on the Hilbert transform and a new approach based on solving a convex optimization problem. We show that convex demodulation better separates the speech information between the modulator and the carrier. We demonstrate this advantage by measuring the speech-information content using a speech-recognition experiment. Finally, we explore the effect of subband filtering on the demodulation process and the shift of information from the modulator to the carrier as the subbands become wider.

Index Terms— Modulation, Optimization methods, Speech recognition, Hilbert transforms

1. INTRODUCTION

There is renewed interest in the fine-temporal structure of speech signals decoded by the auditory system[1]. Researchers would like to know if the human auditory system is using the temporal information in a cochlear channel to help decode speech. This has become especially important for producers of cochlear implants, which have a limited number of channels, yet want to deliver the maximum amount of information to the wearer. This paper reports a decomposition of the signal into modulators and carriers at several channel bandwidths and describes the information content of each with a speech recognition task

Many experiments studying the fine-time structure of speech signals use a Hilbert transform to decompose the signal. They further assume the "envelope cues carry most of the information required for speech identification... with [the carrier] primarily conveying pitch cues."[2] The Hilbert transform, which is based on the analytic signal, is a perfect demodulator in the simplest case (a single sinusoidal carrier with a low frequency modulator). But, for more challenging cases, such as carriers with multiple sinusoids, the Hilbert envelope mixes carrier and modulator content, resulting in intermodulation terms in both signals..

Our demodulation work is based on a new formulation using convex optimization[3]. Demodulation is inherently an ill-posed problem with an infinite set of solutions. To select an optimal solution from this set, we use a cost function to describe the properties of the optimal estimated modulation and carrier. In brief, for the linear-optimization criteria we use in this paper, the optimization method finds a modulator with minimal high-frequency content that most closely fits the amplitude of the original signal.

Our analysis is based on a subband decomposition of the auditory signal. The auditory system performs a rough spectral analysis using the cochlea, and the inner-hair cells transduce different spectral bands of the signal into neural firings. We show the information content of the estimated modulator and carrier using two different demodulation algorithms, and show how the information changes with channel bandwidth.

We do not argue that the brain uses optimization theory to separate the modulation from the carrier. We use convex optimization to do the decomposition so we can better understand the information content in the two signals. In Section 3, we describe a simple vowel-recognition experiment to demonstrate the improvement that can be achieved with convex demodulation.

2. CHANNEL DEMODULATION

Signal demodulation, at the simplest level, is the decomposition of a signal $s(t)$ in a low-frequency modulator $m(t)$ and a high-frequency carrier $c(t)$

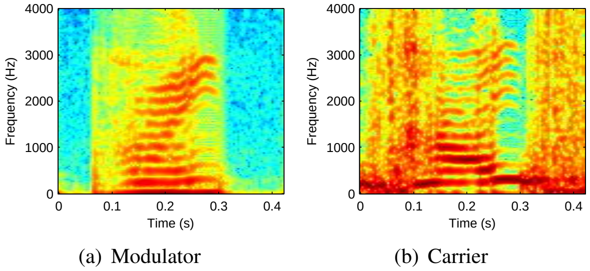
$$s(t) = m(t)c(t).$$

However, it is also possible to pose the problem as a sum of modulated sources[4]

$$s(t) = \sum_k s_k(t) = \sum_k m_k(t)c_k(t).$$

These sources could be defined in any number of ways, but for most applications, such as modulation filtering[5] and chimaeric speech perception studies[1, 6], $s_k(t)$ is defined as a subband channel.

*Gregory Sell was supported by the Ric Weiland Graduate Fellowship



(a) Modulator (b) Carrier

Fig. 1. Plots of the wideband Hilbert demodulated components for the word "pipe" as spoken by a female speaker. The speech information, both pitch and formants, is easily seen in both signals.

2.1. Mixing of Speech Information in a Hilbert Modulator and Carrier

The Hilbert envelope, which is defined as the magnitude of the analytic signal obtained from the Hilbert Transform, is widely used in demodulation experiments. However, in the case of wideband modulation, the Hilbert envelope does not properly separate the speech information between the modulator and the carrier. In fact, the information in the Hilbert envelope of a speech signal is easily recognizable to a human listener.

To demonstrate this visually, Fig. 1 shows the spectrogram of the wideband Hilbert envelope and carrier derived from the spoken word "pipe." Both signals clearly show pitch and formant information. This indicates that the Hilbert envelope is encoding speech information that a modulator should not contain, based on the low-frequency constraint.

This failure of the Hilbert envelope demonstrates that the method is not sufficient for audio demodulation tasks, especially those that examine the presence of information in the modulator and carrier, such as chimaeric speech experiments. We believe that convex demodulation is more effective for these tasks.

2.2. Convex Demodulation

Instead, we pose the demodulation task as an optimization problem[3]. In the linear-domain method used in this paper, the optimization problem uses a cost function that minimizes the presence of high frequencies using a spectral penalty $W(f)$, which is a sigmoidal function in the frequency domain, as well as minimizing the norm of the modulator. The optimal modulator is then found by minimizing the cost function subject to signal-dependent amplitude constraints designed to ensure the modulator will closely match the envelope of the original signal. The algorithm specifically solves the optimization problem:

$$\begin{array}{ll} \text{minimize} & ||W(f)\mathcal{F}\{m(t)\}||_2^2 + ||m(t)||_2^2 \\ \text{subject to} & m(t) - 1 \leq 0, \forall t \\ & |s(t)| - m(t) \leq 0, \forall t. \end{array}$$

Note that $s(t)$ can be the entire wideband signal, or a single subband channel. Once the optimization problem is solved, the convex carrier then follows by dividing the original signal by the derived modulator.

The advantage of this method over the Hilbert envelope is that it is specifically designed to solve for a low frequency modulator for any carrier, including the harmonic and stochastic carriers that result in problematic decompositions with the Hilbert envelope.

3. SPEECH RECOGNITION AFTER CHANNEL DEMODULATION

To demonstrate the improved performance that convex demodulation offers over the Hilbert envelope for audio tasks, we designed a speech-recognition experiment to measure the information content in the two output signals. We used synthesized vowels with random pitch and added noise. We performed the demodulation and then tested to see how much of the formant information remained. Ideally, the speech (formant) information should be in one signal or the other.

3.1. Methods

The synthetic voice signals were created with the Auditory Toolbox[7] using the vowels /a/, /i/, and /u/ at a random pitch between 120 and 300 Hz. Each signal has a random amount of noise, resulting in an SNR of at least 10 dB, and 6000 utterances were used for each channel bandwidth.

Each experiment consisted of decomposing the vowel utterances with a filterbank of a certain channel bandwidth. Each channel was then demodulated using both Hilbert envelope extraction and convex demodulation. The channel modulators and carriers for each case were used to create wideband components.

Wideband carriers were reconstructed by summing the extracted channel carriers. Wideband modulators were created by multiplying each estimated channel modulator by a sinusoid at the channel's center frequency and then summing. This was used instead of filtered noise because noise is potentially destructive to the spectral content, and the classification task we chose only uses the smoothed spectral shape, and so the sinusoidal carriers in modulator reconstruction did not play a significant role beyond shifting the modulator spectrum to the correct band. Mel Frequency Cepstral Coefficients (MFCC)[7] of the wideband components were used for classification.

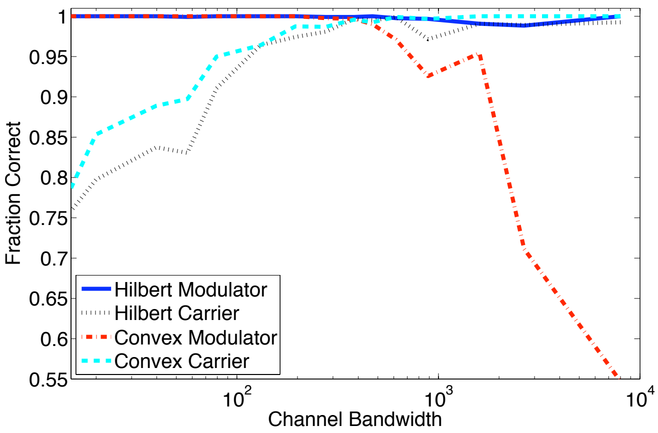


Fig. 2. Results for the speech-recognition experiment described in Section 3. The graph shows the recognition accuracy (and therefore speech information content) for the four signals plotted against the bandwidth of the channel filters.

A k-nearest neighbors classifier was trained in each case on demodulated training signals with the same channel characteristics as the test signals. This was selected rather than using a single classifier trained on real speech because we want to determine if there is any information present in the modulator or carrier, but not necessarily to see if this information matches up with the original speech signal well enough for proper classification. A majority decision based on the five nearest neighbors determined the classification result.

Signal processing concerns when combining adjacent channels are not a problem for two reasons. First, the same processing is applied to all signals. Since the purpose of these experiments is to compare the signals, the artifacts are consistent across conditions and relative comparisons are valid. Second, as stated above, the task is based only on identifying the presence of speech information, not on creating intelligible audio signals, so distortion of intelligibility to a human listener is not an issue for this specific task.

This methodology offers the best possible conditions for Hilbert envelope detection, in that only vowels are used. Consonants would only provide a greater challenge, and nonharmonic carriers, as is the case for whispering, for example, are more problematic for Hilbert envelope detection.

3.2. Results

Fig. 2 shows the recognition accuracy and thus the information content of the signals at different channel bandwidths for both demodulation methods. In the convex case, the modulator yields low accuracy for wideband channel demodulation (right side) and high accuracy for narrowband demodulation (left side). The carrier shows the opposite performance. Between the two extremes, both signals yield moderate recognition, indicating both have information. The X shape of the two convex curves shows that the speech information shifts from the carrier to the modulator as the bandwidth increases.

For the Hilbert signals, the results show the modulator contains the speech information across all bandwidths, while the speech information in the Hilbert carrier follows a similar trajectory to the convex case, decreasing its accuracy with channel bandwidth. Unless a narrowband channel is used, the speech information is present in both the modulator and carrier, demonstrating the inability of the Hilbert transform to properly separate the two.

While we have built a very strong classifier for this task, we can only argue that these performance curves represent a lower bound on the information present in the signal. We have used MFCC features, just like the state-of-the-art speech-recognition systems. We trained and tested the classifier on data with identical bandwidth and noise levels. Yet it is possible that the demodulation algorithms we have used have put the vowel information somewhere in the signal that MFCC features do not see. Humans might do better at this simple task then machines do.

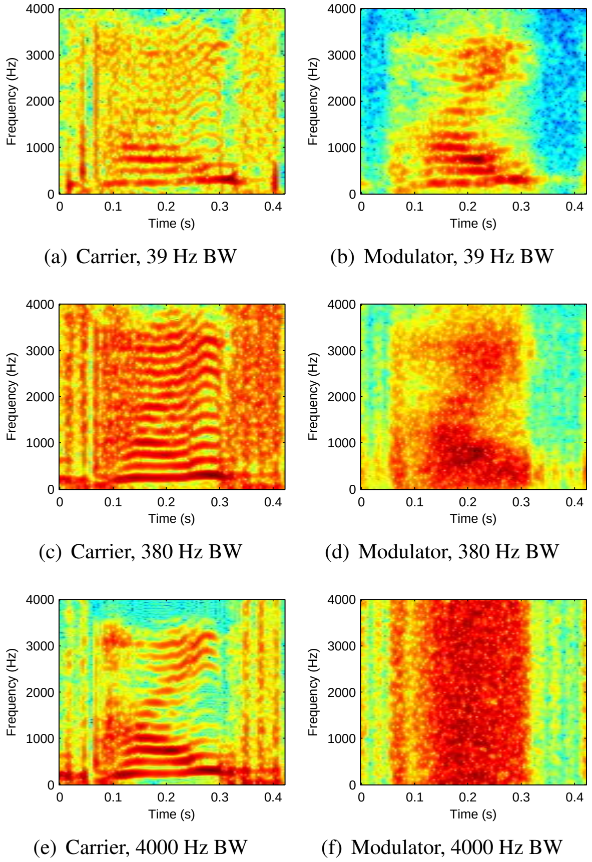
It is also worth noting that our curve for the Hilbert modulator is significantly different for wideband channels than the results seen in tests performed by Smith et al [6]. This is because the two experiments tested different aspects of information content. In the experiments performed by Smith et al, the intent was to find the dominant information content between the modulator and carrier. Our test determines if information is present at all in either the modulator or carrier, and the experiment clearly shows that the Hilbert modulator does have speech information in wideband channels.

3.3. Discussion

Fig. 2 shows the shift in information from the carrier at narrow bandwidths to the modulator at wide bandwidths. This is an interesting effect, and the reason for it is directly related to the bandwidth of the channels, and its relation to the spectral spacing of formants in speech.

Depending on the bandwidth of the subband decomposition, demodulation puts different information in the carrier and modulator. Channel demodulation, when done properly as with convex demodulation, yields a low-frequency modulator that only represents the time-varying amplitude envelope of the channel. The carrier holds all of the relative amplitudes of the harmonics within the channel. The modulator only contains the overall amplitude of the channel. So, when the carriers and modulators from the channels are recombined, the relative spectral amplitude differences within the channels are still present in the carrier. However, relative amplitude differences between the channels are removed from the carrier by the demodulation, and the modulator contains any relevant information that is encoded by those differences.

Therefore, information that is found with wider spectral



(a) Carrier, 39 Hz BW (b) Modulator, 39 Hz BW

(c) Carrier, 380 Hz BW (d) Modulator, 380 Hz BW

(e) Carrier, 4000 Hz BW (f) Modulator, 4000 Hz BW

Fig. 3. Spectrograms of the convex demodulation components of the word "pipe" at three channel bandwidths.

spacing than the channel bandwidth is placed in the modulator, because the relative amplitudes of the channels can encode it. Information that is found with a narrower spectral spacing than the channel bandwidth is placed in the carrier.

We can see this in the convex signal performances in Fig. 2, where the speech information moves between the carrier and the modulator in the range of 300-700 Hz, roughly. This is a reasonable range for the spacing of the formant frequencies in the vowels used for the experiment. So, the speech information moves from the modulator to the carrier as the channel bandwidth increases into and then past the formant spacing. The modulator performance does not drop until the channel bandwidth exceeds 2000 Hz, which is roughly the widest formant spacing in the vowels used.

Fig. 3 illustrates this tradeoff between formant information in the carrier versus the modulator. This figure shows the modulators and carriers determined at three channel bandwidths. With narrowband channels, the formant and even pitch data can be seen in the modulator in Fig. 3(b), while the carrier in Fig. 3(a) is missing most of that information. In a mid-range bandwidth, the formant data is clearly present in the modulator in Fig. 3(d) while the pitch is in the carrier in Fig. 3(c). For a wideband channel, the carrier in Fig. 3(e) has all of the relevant speech information while the modulator in Fig. 3(f) has only the overall amplitude of the signal, containing no speech information beyond the plosive consonant 'p' at the beginning of the file.

This example is extremely relevant to the demodulation problem, because it clearly shows that the definition of modulation and the information its components contain is highly dependent on the channel filter bandwidth.

4. CONCLUSION

Identifying the speech and voicing information via demodulation of a signal is an important task for improving performance of cochear implants and hearing aids, among other applications. We presented results demonstrating that convex demodulation separates speech information between modulators and carriers, even for wideband channels, where Hilbert does not separate the information. This advantage is evident for the simplest speech-like test signals, stationary vowels. Properly separating the demodulation components is essential for rigorous testing of the location of speech information, and so it is our opinion that convex demodulation offers higher quality decompositions for future experiments testing the presence of speech information in demodulated signals.

5. REFERENCES

[1] M. G. Heinz and J. Swaminathan, "Quantifying Envelope and Fine-Structure Coding in Auditory-Nerve Responses to Chimaeric Speech," *JARO*, vol. 10, no. 3, pp. 407–423, Sept. 2009.

[2] S. Sheft, M. Ardoint, and C. Lorenzi, "Speech identification based on temporal fine structure cues," *JASA*, vol. 124, no. 1, pp. 562–575, July 2008.

[3] G. Sell and M. Slaney, "Solving Demodulation as an Optimization Problem," Accepted for publication, IEEE TASLP., 2010.

[4] P. Clark and L. Atlas, "A sum-of-products model for effective coherence modulation filtering," in *Proceedings of the ICASSP*, 2009.

[5] P. Clark and L. Atlas, "Time-Frequency Coherent Modulation Filtering of Nonstationary Signals," *IEEE TSP*, vol. 57, no. 11, pp. 4323–4332, Nov. 2009.

[6] Z. M. Smith, B. Delgutte, and A. J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Nature*, vol. 416, no. 6876, pp. 87–90, March 2002.

[7] M. Slaney, "Auditory Toolbox," Tech. Rep. 1998-010, Interval Research Corporation, 1998.