FEASIBILITY OF SINGLE CHANNEL SPEAKER SEPARATION BASED ON MODULATION FREQUENCY ANALYSIS

Steven M. Schimmel, Les E. Atlas* and Kaibao Nie***

*Department of Electrical Engineering, **VM Bloedel Hearing Research Center,
University of Washington, Seattle, WA 98195, USA

ABSTRACT

We explore the use of the modulation frequency domain for single channel speaker separation. We discuss features of the modulation spectrogram of speech signals that suggest that multiple speakers are highly separable in this space. In a preliminary experiment, we separate a target speaker from an interfering speaker by manually masking out modulation spectral features of the interferer. We extend this experiment into a new automatic speaker separation algorithm, and show that it achieves an acceptable level of separation. The new algorithm only needs a rough estimate of the target speaker's pitch range.

Index Terms — Speech enhancement, separation, modulation, spectral analysis, time-varying filters

1. INTRODUCTION

A common complaint among users of cochlear implants and hearing aids is the inability to focus on a single speaker in situations with multiple interfering speakers, such as in bars, restaurants and other places of social gathering. A study by Bronkhorst and Plomp [1] showed that under those conditions the hearing impaired need 4–10 dB better SNR than the normal hearing for equal intelligibility.

Modern hearing instruments address this "cocktail party" problem in several ways [2]. They use directional microphones or microphone arrays to enhance speech from the front, thus reducing interference from speakers from other directions, and adaptive (single channel) noise suppression techniques to reduce background noise. Despite the fact that these methods improve the SNR and can reduce the listening stress, they have yet to prove that they enhance speech intelligibility [2].

Beyond the context of hearing aids, several approaches to solve the single channel cocktail party problem have been proposed. Two prominent classical techniques are adaptive comb filtering [3] and harmonic selection [4]. Among the more recent techniques are harmonic enhancement and suppression [5] and pitch tracking and amplitude modulation [6]. All these techniques, however, require very accurate pitch estimates, which is a difficult problem in itself for single speakers, and even more so in the presence of interfering speakers.

In this paper, we propose to approach the single channel cocktail party problem in the *modulation frequency domain*. This choice is motivated on one side by psychoacoustic research, e.g. the work by Dau *et al.* [7], that support the belief that the human auditory system also analyzes and possibly even segregates sounds in this domain [8]. The modulation frequency domain has been used before as a new approach to existing problems, for example for speaker recognition [9], and automatic speech recognition [10]. It has also been used, by Kollmeier and Koch [8], to address the cocktail party problem in two channels. They used phase and intensity differences between modulation frequency representations of stereo channels to separate speakers. The objective of this paper, however, is to show how features of the modulation frequency representation of a single channel of speech from multiple speakers can be used for speaker separation.

Another argument in favor of a modulation frequency domain approach is that it only requires a rough estimate of a desired speaker's pitch range and that it takes only a simple algorithm to achieve an acceptable level of speaker separation, as we will demonstrate in this paper.

The paper is organized as follows. Section 2 gives a general introduction of modulation frequency analysis and the modulation spectrogram. Section 3 describes a preliminary manual experiment and its results that demonstrate the modulation spectral features of interest, and their usability for speaker separation. Section 4 discusses a new method of automatic speaker separation that is based on the results of the manual separation experiment. Finally, section 5 presents the speaker separation results obtained with the automatic separation technique, followed by conclusions and a discussion in section 6.

2. MODULATION FREQUENCY ANALYSIS

The general modulation frequency analysis framework consists of a filterbank (possibly decimated), followed by subband envelope detection and frequency analysis of the subband envelopes. In its most straightforward form, the filterbank is implemented using the short-time Fourier transform (STFT), envelope detection is defined as the magnitude or magnitude squared of the subband, and subband envelope frequency analysis is performed with the Fourier transform. For a discrete signal $x(n)$, the STFT can be expressed as

$$X_k(m) = \sum_{n=-\infty}^{\infty} h(mM - n)x(n)W_K^{kn}, \quad \text{for } k = 0,\ldots,K-1, \tag{1}$$

and the envelope detection and modulation frequency analysis as

$$X_l(k,i) = \sum_{m=-\infty}^{\infty} g(lL - m)\left|X_k(m)\right|W_I^{im}, \quad \text{for } i = 0,\ldots,I-1, \tag{2}$$

where $W_K = e^{-j(2\pi/K)}$. $h(n)$ and $g(m)$ are the acoustic and modulation frequency analysis windows, respectively. Throughout this paper we will use the shorthand notations

$$T\{x(n)\} = X_l(k,i) \tag{3}$$

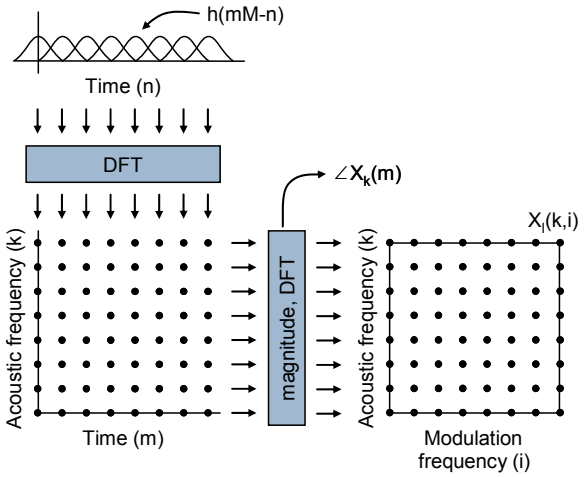

Figure 1 Modulation analysis framework and the modulation spectrogram.

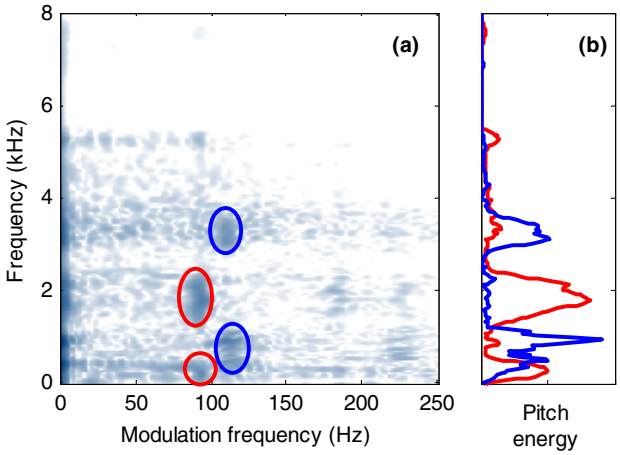

Figure 2 (a) Modulation spectrogram of the sum of two speakers. Highlighted features are the pitch energy of speaker A (*red*) and speaker B (*blue*). **(b)** Acoustic frequency localization of pitch energy of speaker A (*red*) and B (*blue*).

and

$$T^{-1}\{X_l(k,i)\} = x(n) \tag{4}$$

to refer to modulation frequency analysis and synthesis.

The magnitude of the subband envelope spectra $|X_l(k,i)|$ is typically displayed in a *modulation spectrogram* representation. The vertical axis of this representation is regular acoustic frequency (k), and its horizontal axis is modulation frequency (i). Gray-scale intensity or color in the joint acoustic/modulation plane represents modulation spectral energy. The modulation analysis framework is illustrated in Figure 1, and an example of a modulation spectrogram is shown in Figure 2a.

There are two types of modulation frequency analyses, wideband and narrowband, which are controlled by the length of the analysis windows $h(n)$. When $h(n)$ is short (wideband analysis), the frequency subbands will be wide and the maximum observable modulation frequency is high. When $h(n)$ is long (narrowband analysis), the frequency subbands will be narrow and the maximum observable modulation frequency is low. For our application we only consider wideband modulation frequency analysis. We want the maximum modulation frequency to be around 300 Hz, so that an adult speaker's pitch resolves in modulation frequency.

The experiment described in the following section demonstrates how a speaker's pitch in modulation frequency can be used to localize a speaker in acoustic frequency.

3. EXPERIMENT: MANUAL SPEAKER SEPARATION

Modulation spectrograms of speech from multiple speakers exhibit several prominent features that could possibly be exploited to separate speakers. For example, Figure 2a shows the modulation spectrogram of the sum of two speakers, $T\{x_a(n)+x_b(n)\}$, for speaker signals $x_a(n)$ and $x_b(n)$ taken from Te-Won Lee's "a real cocktail party effect" dataset [11]. The figure shows that the modulation spectral energy of the two speakers is concentrated at different modulation frequencies due to their different fundamental frequencies. Moreover, the energy at each speaker's pitch in the modulation frequency dimension is localized in acoustic frequency, and peaks at major acoustic features of its source such as formants, see Figure 2b. The separation of the two speakers in modulation frequency, together with the acoustic frequency localization of pitch energy, suggests that the speakers might be separable in the modulation domain.

To test this hypothesis, we analyzed the modulation spectral content of 262 ms long frames of $x_a(n)$, $x_b(n)$ and their 0 dB mix $x(n) = x_a(n) + x_b(n)$, at 64 ms intervals. Based on the apparent separation of the speakers in the modulation domain, we manually constructed two binary modulation spectral masks, $M_l^a(k,i)$ and $M_l^b(k,i)$, such that

$$\tilde{x}_a(n) = T^{-1}\left\{M_l^a(k,i)\cdot T\{x(n)\}\right\} \approx x_a(n) \tag{5}$$

and

$$\tilde{x}_b(n) = T^{-1}\left\{M_l^b(k,i)\cdot T\{x(n)\}\right\} \approx x_b(n) \tag{6}$$

Figure 3 shows an example of both masks for one frame ($l = 19$) of the signal $x(n)$. We reconstructed time-domain signals from the masked modulation spectrograms by reversing the steps in Eq. (2), and using the original STFT phase

$$\angle X_k(m) = X_k(m)/|X_k(m)| \tag{7}$$

to invert the STFT in Eq. (1). Although reconstruction from a modified STFT introduces distortion, this technique was found to achieve reasonably good signal quality [12]. The result of this experiment is illustrated by the spectrograms in Figure 4.

As the spectrograms in Figure 4 show, formants are well preserved and assigned to the correct speaker, and most voiceless sounds are also separated well. The loss of low frequencies in the voiced fricative /s/ in "dos" and "tres" is audible as slight amplitude pumping. Although the spectrograms don't show it, there's also some audible crosstalk between the talkers at word onsets and offsets. The separation of the speakers is good, but the non-linearity of the modulation domain signal processing has added distortion to the signals [13], giving them a slight metallic or tinny quality.

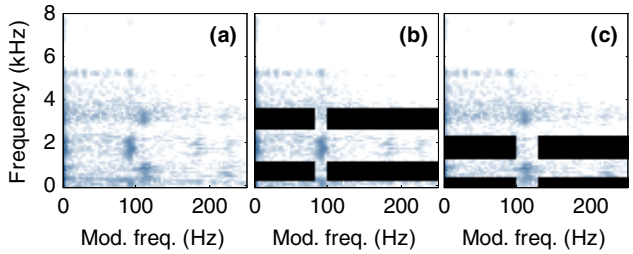

Figure 4 (a) Modulation spectrogram $T\{x_a(n)+x_b(n)\}$ **(b)** Mask $M_{19}^{a}(k,i)$ **(c)** Mask $M_{19}^{b}(k,i)$. Black indicates regions that will be masked out (i.e. set to zero) during signal reconstruction.

4. METHOD: AUTOMATIC SPEAKER SEPARATION

To demonstrate the feasibility of using a speech signal's modulation spectral representation to automatically separate two speakers, we devised the following separation algorithm. Given a signal sampled at f_s Hz that is the sum of a target speaker and an interfering speaker, i.e. $x(n)=x_t(n)+x_i(n)$. Assume that the target speaker's pitch is within the fixed frequency range $P_t=[f_{t,low},f_{t,high}]$, and that the interfering speaker's pitch is within the range $P_i=[f_{i,low},f_{i,high}]$. The pitch ranges can be broad, but should be sufficiently non-overlapping. Define $Q=\{\, i : i(f_s/IM)\in P\}$ as the set of modulation frequency indexes *i* in the pitch range P, and let $X_l(k,i)=T\{x(n)\}$. Consider the modulation spectral energy as a function of acoustic frequency index over the target's pitch range

$$E_l^t(k)=\sum_{i\in Q_t}\left|X_l(k,i)\right|^2, \tag{8}$$

as well as over the interfering speaker's pitch range

$$E_l^i(k)=\sum_{i\in Q_i}\left|X_l(k,i)\right|^2. \tag{9}$$

For each frame *l*, i.e. at the time instances $n=l(LM)$, target energy and interferer energy define a frequency masking function

$$F_l(k)=\frac{E_l^t(k)}{E_l^t(k)+E_l^i(k)}. \tag{10}$$

To mask out the interfering speaker and reconstruct a time-domain signal, we decided not to use modulation filtering and modulation synthesis, to avoid the known artifacts associated with it [13-15]. Instead, each frame's frequency masking function is transformed to an impulse response by combining it with the appropriate phase response $\phi(k)$ and taking the inverse DFT,

$$f_l(n)=\frac{1}{N}\sum_{k=0}^{N-1}F_l(k)\phi(k)\,W_N^{-kn}. \tag{11}$$

A time-varying filter $h_k(n)$ is constructed for all times *k* by linear combination of the two nearest impulse responses, according to

$$h_k(n)=(1-\alpha_k)f_{\beta_k}(n)+\alpha_k f_{1+\beta_k}(n), \tag{12}$$

where $\alpha_k=k/LM-\beta_k$, $\beta_k=\lfloor k/LM\rfloor$ and $\lfloor x\rfloor$ is the largest integer smaller than or equal to *x*. The time-varying filter is then used to separate the target speaker from the interfering speaker, as follows

$$\tilde{x}_t(n)=\sum_{k=-\infty}^{\infty}x(k)h_k(n). \tag{13}$$

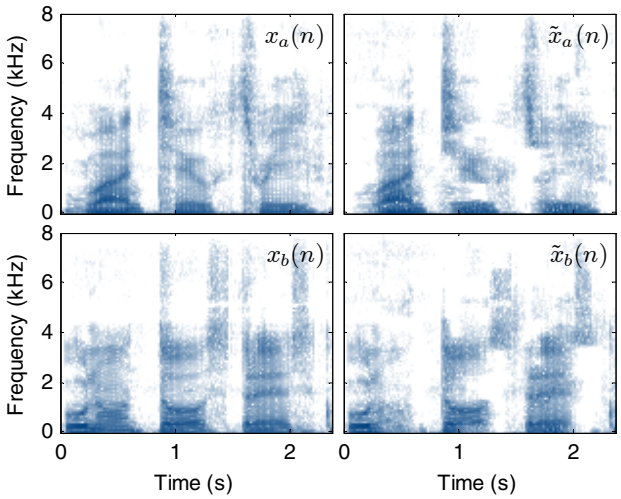

Figure 3 (*top row*) Spectrograms of speaker A ("One, two, three") before and after manual speaker separation.
(*bottom row*) Spectrograms of speaker B ("Uno, dos, tres") before and after manual speaker separation.

5. RESULTS

We selected two stimuli from the TIMIT database, $x_c(n)$ and $x_d(n)$, each spoken by a different male speaker. (Speaker C: "Will Robin wear a yellow lily?" and speaker D: "Do they allow atheists in church?") The stimuli were sampled at 16 kHz. Using a simplified version of a cepstral pitch determination algorithm [16], we estimated the speaker's pitch ranges to be $P_c=[100,124]$ and $P_d=[125,164]$. The algorithm parameters were set to $M=16$, $K=512$, $L=38$, $I=512$, and *h*(*n*) and *g*(*m*) were a 48-point and 78-point Hanning window.

We applied the method described in section 4 to the 0 dB mix of the two stimuli, $x(n)=x_c(n)+x_d(n)$, once with speaker C as the target speaker and speaker D as the interfering speaker, and once with their roles reversed. The result of the automatic separation is illustrated by the spectrograms in Figure 5.

As the spectrograms in Figure 5 show, the automatic separation algorithm allocates the formants (and hence the vowels and other sonorant sounds) to the correct speaker. There is some crosstalk between the speakers, most noticeably the voiceless fricative /s/ in "atheist" at $t=1.3$ and the voiceless affricate /ch/ in "church" at $t=1.7$. The stop /t/ in "atheist" is diminished in the output of speaker D, and the glide /w/ in "wear" is lost in the output of speaker C. Besides these issues with unvoiced sounds, the separated output sounds natural and undistorted, and with reasonably good separation between the speakers.

6. CONCLUSIONS AND DISCUSSION

We demonstrated that acoustic frequency localization of pitch energy is an important feature of speech modulation spectrograms

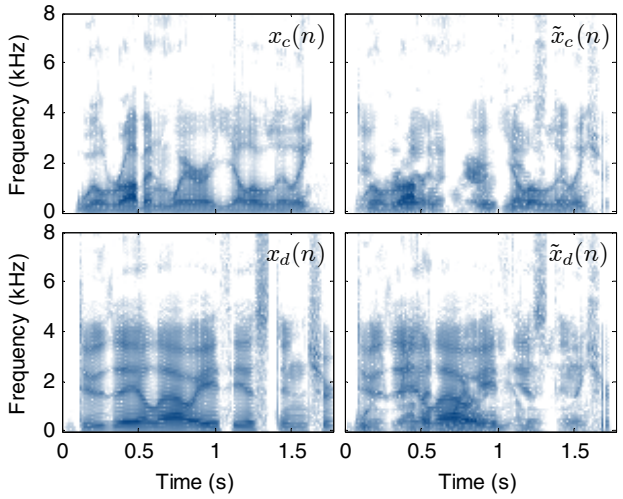

Figure 5 (*top row*) Spectrograms of speaker C ("Will Robin wear a yellow lily") before and after automatic speaker separation. (*bottom row*) Spectrograms of speaker D ("Do they allow atheists in church") before and after automatic speaker separation.

that can be exploited for single channel speaker separation. We presented a new approach for speaker separation based on modulation frequency analysis and a time-varying filter. The proposed method is purposely simple, and designed to require only a rough estimate of pitch, to show that it is possible to achieve a reasonable quality of separation in the modulation spectral domain with relatively simple means. The method is complementary to other approaches, and can be combined with existing techniques such as directional microphones and adaptive noise suppression for increased benefit.

There are additional modulation spectral features that can be exploited for speaker separation. For example, the phenomenon of subband comodulation at low modulation frequencies (2–16 Hz) could be another cue for frequency localization of the target speaker. Furthermore, the presence of pitch energy for the target speaker could be used to keep track of the voiced/unvoiced/silent state of the target speaker, and would enable the algorithm to suppress interfering voiceless speech at times of voiced speech from the target speaker. Finally, the algorithm's pitch model could be extended from a fixed pitch range to, for example, an adaptive pitch range. The adaptive pitch range would follow the target speaker's pitch in modulation frequency, and would enable the algorithm to better separate the target speaker from interfering speakers with similar pitch ranges.

7. ACKNOWLEDGEMENTS

The authors wish to thank Jeffrey K. Thompson of Neural Audio, Kirkland, WA for his work on the manual speaker separation experiment.

8. REFERENCES

[1] A. W. Bronkhorst and R. Plomp, "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing," Journal of the Acoustical Society of America, vol. 92, pp. 3132– 3139, 1992.

[2] J. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass, "Signal processing in high-end hearing aids: state of the art, challenges, and future trends," EURASIP Journal on Applied Signal Processing, vol. 18, pp. 2915–2929, 2005.

[3] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," Proceedings of the IEEE, vol. 67, pp. 1586–1604, 1979.

[4] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," Journal of the Acoustical Society of America, vol. 60, pp. 911–918, 1976.

[5] D. P. Morgan, E. B. George, L. T. Lee, and S. M. Kay, "Cochannel speaker separation by harmonic enhancement and suppression," IEEE Transactions on Speech and Audio Processing, vol. 5, pp. 407–424, 1997.

[6] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," IEEE Transactions on Neural Networks, vol. 15, pp. 1135–1150, 2004.

[7] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system: I. Model structure," Journal of the Acoustical Society of America, vol. 99, pp. 3615–3622, 1997.

[8] B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," Journal of the Acoustical Society of America, vol. 95, pp. 1593–1602, 1994.

[9] T. Kinnnunen, "Joint acoustic-modulation frequency for speaker recognition," Proceedings of ICASSP, pp. 665-668, 2006.

[10] H. Hermansky, "The modulation spectrum in the automatic recognition of speech," Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 140-147, 1997.

[11] T.-W. Lee, "Blind Source Separation: Audio Examples," http://www.snl.salk.edu/~tewon/Blind/blind_audio.htmlT, 1998.

[12] S. M. Schimmel and L. E. Atlas, "Analysis of signal reconstruction after modulation filtering," Proceedings of SPIE, Vol. 5910, pp. 59100H 1–10, 2005.

[13] S. M. Schimmel and L. E. Atlas, "Coherent envelope detection for modulation filtering of speech," Proceedings of ICASSP, pp. 221–224, 2005.

[14] L. Atlas, Q. Li, and J. Thompson, "Homomorphic modulation spectra," Proceedings of ICASSP, pp. 761–764, 2004.

[15] S. M. Schimmel, K. R. Fitz, and L. E. Atlas, "Frequency Reassignment for Coherent Modulation Filtering," Proceedings of ICASSP, pp. 261–264, 2006.

[16] A. Noll, "Cepstrum pitch determination," Journal of the Acoustical Society of America, vol. 41, pp. 293–309, 1967.