CO-CHANNEL SPEAKER SEPARATION USING CONSTRAINED NONLINEAR OPTIMIZATION

Daniel S. Benincasa
Rome Laboratory/OCSS
26 Electronics Pkwy
Rome, NY 13441-4514, USA

Michael I. Savic
ECSE Department/Speech Research Group
Rensselaer Polytechnic Institute
Troy, NY 12180, USA

ABSTRACT

This paper describes a technique to separate the speech of two speakers recorded over a single channel. The main focus of this research is to separate overlapping *voiced* speech signals using constrained nonlinear optimization. Based on the assumption that voiced speech can be modeled as a slowly-varying vocal tract filter with a quasi-periodic train of impulses, the speech waveform is represented as a sum of sine waves with time-varying amplitude, frequency and phase. In this work the unknown parameters of our speech model will be the amplitude, frequency and phase of the harmonics of both speech signals. Using constrained nonlinear optimization, we will determine, on a frame by frame basis, the best possible parameters that provides the least mean square error (LMSE) between the original co-channel speech signal and the sum of the reconstructed speech signals.

1. INTRODUCTION

In many situations, the intelligibility of a person's speech, recorded over a single channel, can be significantly degraded due to the linear addition of speech from persons other than the desired speaker. An example of such a situation is the speech of the airline pilot recorded from the cockpit of an aircraft in distress. Co-channel speaker separation is the process by which the speech of the desired speaker (pilot) is extracted from the co-channel speech (sum of the speech from the pilot and copilot). In this paper, we will address the case of the overlapping speech of two speakers.

Previous authors attempting to solve the co-channel speaker separation problem have presented techniques which require the use of *a priori* information [1,2], typically assuming the position, amplitude or phase of the true spectral harmonics are either available or not required. Still others have presented techniques which estimate the spectral magnitude of a desired speech signal through harmonic suppression using spectral magnitude subtraction of the interfering signal from the co-channel signal [3,4]. Our research is unique in that it looks to optimize all three parameters, frequency, phase and amplitude for the harmonics of both speakers.

This paper is organized as follows. Section two will discuss the sinusoidal model for speech. Section three will address constrained nonlinear optimization. Section four will describe the speaker separation system and section five will provide preliminary results with speech signals.

2. SINUSOIDAL MODEL

A short segment of voiced speech can be modeled as a slowly-varying vocal tract filter with a quasi-periodic train of impulses as the driving source. The speech waveform can be represented as a sum of sine waves with time-varying amplitude, frequency and phase [5]. We can simplify this model by assuming the excitation signal is quasi-periodic. Then each sinusoidal component can be represented with a fixed frequency, amplitude and phase. Our speech signal, *s(n)*, can then be written as

$$s(n) = \sum_{k=1}^{M} A_k \cos\left[2\pi n f_k + \phi_k\right] \qquad (1.1)$$

where the amplitude, frequency and phase are denoted by A_k, f_k, and ϕ_k respectively summed over M harmonics. Using this model, we can reconstruct the speech waveform by sampling the spectrum at the harmonic peaks to obtain values for the amplitude, frequency and phase.

The spectrum of the co-channel speech may contain some of the harmonics of each speech signal. We use this spectrum to provide an initial guess to A_k, f_k, and ϕ_k for both speech signals and then impose constrained nonlinear least squares optimization to determine the best possible parameters of the two speech signals which resulted in the co-channel speech.

3. OPTIMIZATION

Optimization involves finding the best solution to a particular problem. Mathematically, this means finding the maximum or minimum of some function of n variables. This function will be referred to as an objective function. When dealing with non-linear objective functions, optimal solutions become more difficult with the potential of local

minima posing as erroneous solutions. For the general non-linear programming problem of the form

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && g_i(x) \le b_i \quad (i = 1, ..., m) \\ & && x_j \ge 0 \quad (j = 1, ...n) \end{aligned} \tag{1.2}$$

We wish to replace the objective function $f(x)$ by

$$f(x) + \sum_{i=1}^{m} \pi_i \left(g_i(x) + y_i - b_i \right) - \sum_{j=1}^{n} \delta_j \left(x_j - s_j \right) \tag{1.3}$$

where the π_i and δ_j are called Lagrange multipliers and (1.3) is referred to as the Lagrangian function [6]. Optimization is obtained by minimizing this new objective function with respect to y_j and s_j. This is accomplished using a method known as sequential quadratic programming which follows a similar technique developed for unconstrained optimization, known as the quasi-Newton method. Here we look for a solution x^* which causes our objective function to be a minimum while simultaneously solving our gradients to zero and our Hessian to be positive definite. We can approximate the Hessian of the Lagrangian function using the quasi-Newton updating method. The gradient and the Hessian are then used to form a quadratic programming sub-problem to determine the best search direction. The variables of the objective function are modified based on the best search direction. This iterative procedure is conducted until the objective function is minimized.

4. SPEAKER SEPARATION SYSTEM

Referring to Figure 1, the co-channel speech signal, $S_C = S_D + S_U$, which is the sum of the desired and undesired speech signals, is broken into segments or frames 30 msec. in length. Each frame is processed separately. A pre-processor extracts a set of features $\overline{P}_i$, to predict the voicing state of each speaker. Possible states for a speaker are voiced, unvoiced or silence. Once the voicing state has been determined, each co-channel speech segment is processed accordingly to the particular states present.

When the desired speech is unvoiced and the interfering speech is voiced, the signal is highpass filtered to remove the effects of the undesired speech signal. When the desired speech is voiced and the interfering speech is unvoiced, the co-channel speech is lowpass filtered, again to remove the effects of the undesired speech. Constrained nonlinear optimization is used when both the desired and undesired speech signals are voiced.

For constrained optimization, the co-channel speech follows two separate paths. In the lower path, the co-channel signal passes through a discrete Fourier transform. Initial conditions of the harmonic parameters, for both signals are estimated from this spectrum. These values are then used to provide constraints for the optimization routine which looks at minimizing the least mean square error between the original co-channel speech segment and the reconstructed co-channel speech segment, using the sinusoidal model.

Once a minimum has been determined, and the desired and interfering speech segments have been estimated, an overlap and add technique is used to piece the speech segments into intelligible speech.

5. RESULTS AND DISCUSSION

Tests using the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus have demonstrated excellent results. An example of separating two different vocalic speech segments from two different speakers (male/female) is shown in Figures 2 and 3. The signal was mixed at a signal to interference ratio (SIR) of 0 dB. A plot of the original speech segments, prior to mixing, are shown in Figure 2a and 2b respectively. The co-channel speech segment, plotted in Figure 2c is the sum of these two speech signals. The co-channel speech is passed through the constrained nonlinear optimization branch of the co-channel speaker separation system. The reconstructed speech segments are shown in Figure 3a and 3b respectively. Finally, results using our speaker separation system with two speech signals mixed at a SIR=-6 dB are presented. We assumed the voicing state and pitch of both speakers are known. Results are provided in Figures 4 and 5. Approximately 1/3 of the co-channel speech segments processed were voiced/voiced mixtures. Segments of co-channel speech in which a speaker is in transition (onset or offset of voicing) proved to be the most difficult to separate.

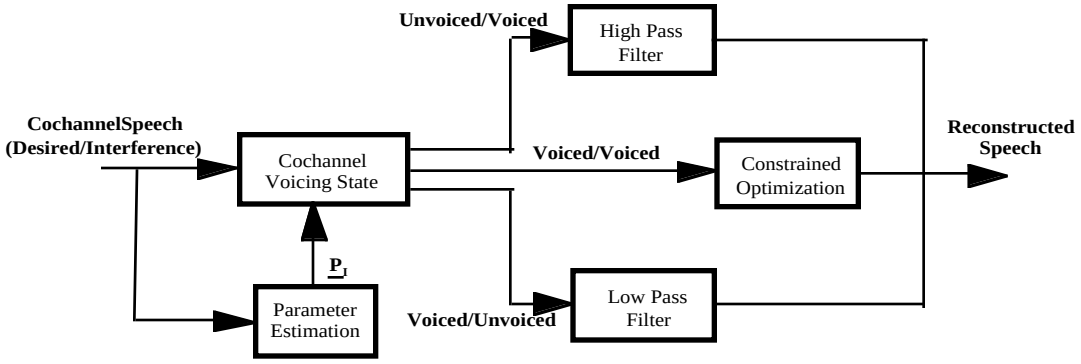
Our optimization algorithm can benefit from improvements in spectral estimation of the harmonic parameters for each speaker. These improvements will provide more accurate constraints, thereby increasing the speed and accuracy of convergence.

REFERENCES

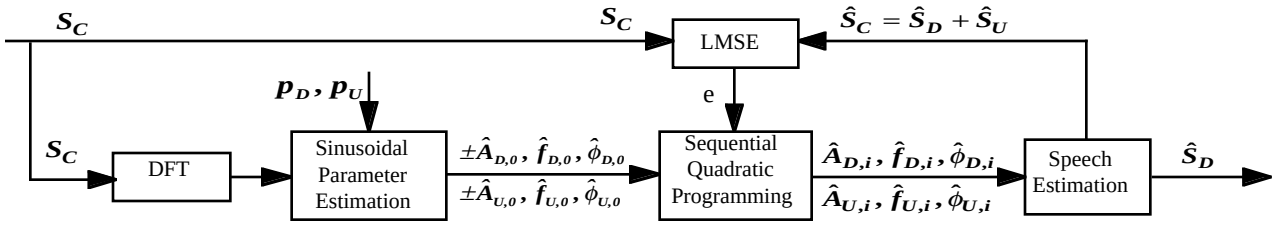
1. R.G. Danisewicz and T.F. Quatieri, "An Approach to Co-channel Talker Interference Suppression Using a Sinusoidal Model for Speech," Technical Report 794, Lincoln Laboratory, MIT, 5 February 1988.
2. D.G. Childers and C.K. Lee, "Co-Channel Speech Separation," Intl. Conf. on Acoust. Speech and Signal Process., Dallas, Texas, April 1987, Vol. 1, pp. 181-184.
3. J. Naylor and S.F. Boll, "Techniques for Suppression of an Interfering Talker in Co-Channel Speech," Intl. Conf. on Acoust. Speech and Signal Process., Dallas, Texas, April 1983, Vol. 1, pp. 205-208.
4. B.A. Hanson and D.Y. Wong, "Processing Techniques for Intelligibility Improvement to Speech with Co-channel Interference," Final Technical Report, Rome Air Development Center (September 1983), DTIC AD-A135702.
5. R.J. McAulay and T.F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal

Representation", IEEE Trans. Acoust. Speech Signal Process., Vol ASSP-34, No. 4, August 1986.

6. M.J.D. Powell, "The Convergence of Variable Metric Methods for Nonlinearly Constrained Optimization Calculations," *Nonlinear Programming 3*, (O.L. Mangasarian, R.R. Meyer and S.M. Robinson, eds.), Academic Press, 1978.



(a) System Diagram



(b) Constrained nonlinear least squares optimization

Figure 1: Block diagram of Speaker Separation System

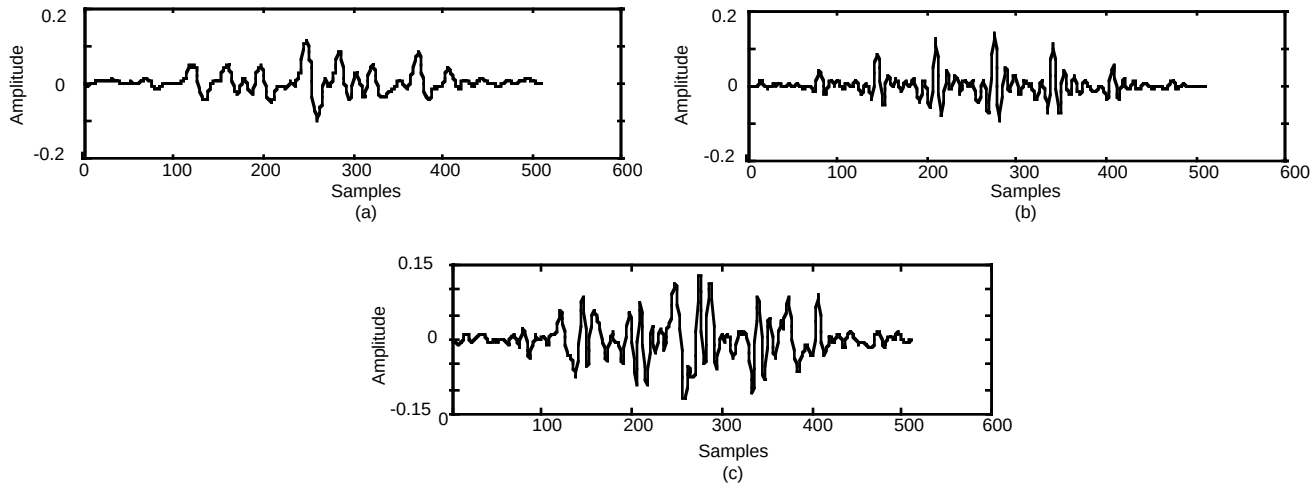


Figure 2: Original Speech Segments: (a) Male speech segment, (b) Female speech segment, (c) Co-channel speech segment

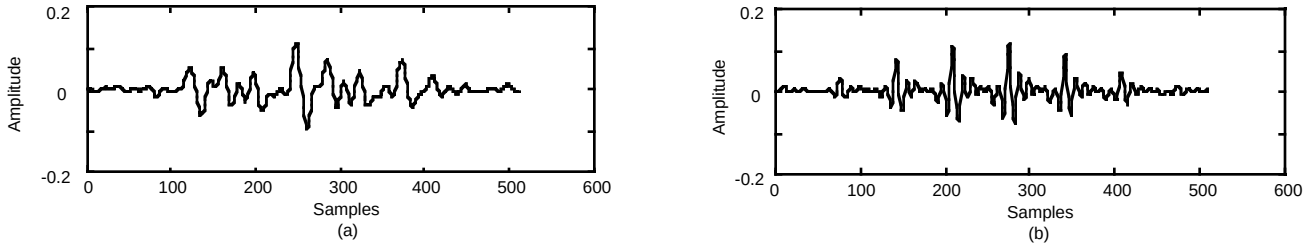


Figure 3: Reconstructed speech segments using optimization: (a) Male speech segment, (b) Female speech segment.

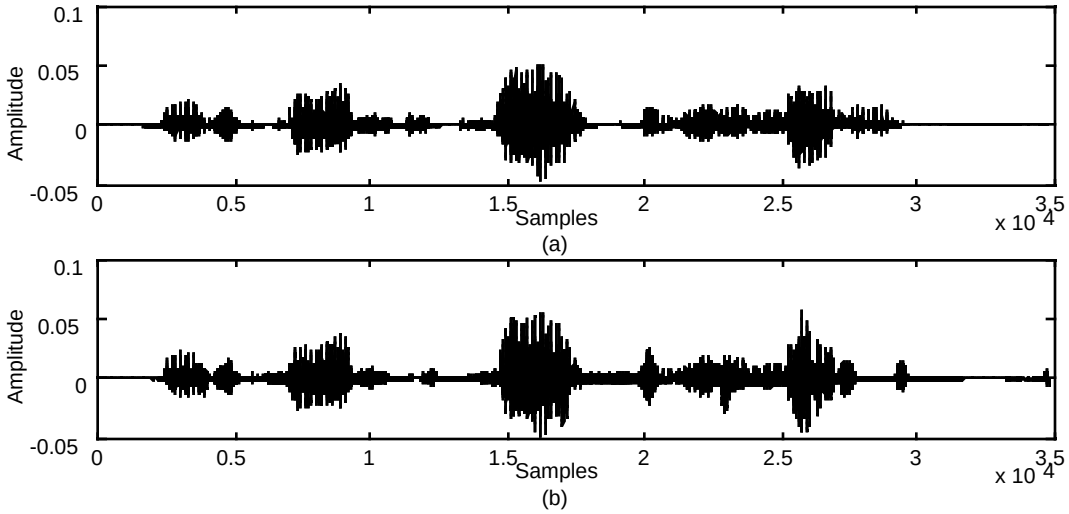


Figure 4: Comparison of original speech and reconstructed speech for the weaker speaker.

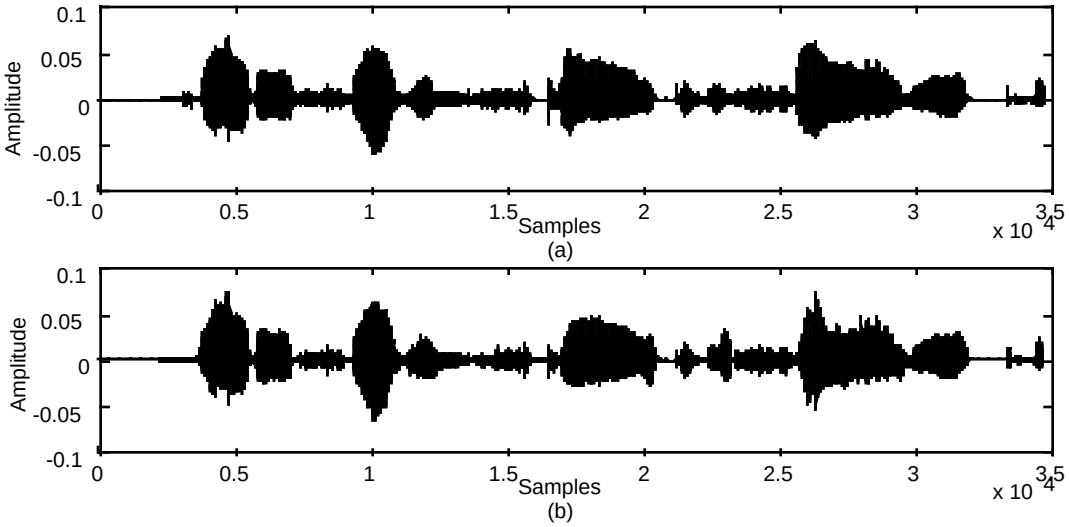


Figure 5: Comparison of original speech and reconstructed speech for the stronger speaker.