Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 28 (2007) 1351–1359
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Foreground auditory scene analysis for hearing aids

Marie A. Roch [a,*], Richard R. Hurtig [b], Tong Huang [a], Jing Liu [a], Sonia M. Arteaga [a]

[a] Department of Computer Science, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182-7720, United States
[b] Department of Speech Pathology and Audiology, The University of Iowa, 119 SHC, Iowa City, IA 52242, United States

Available online 16 March 2007

Abstract

Although a wide variety of signal enhancement algorithms are available for hearing aids, selection and parameterization of the best algorithm at any given time is highly dependent upon the environment of the hearing aid user. The use of auditory scene analysis has been proposed by several groups to categorize the background noise. In this work, an algorithm is proposed to categorize a foreground speaker as opposed to the background noise and parameterize a frequency-based compression algorithm which has been previously shown to improve speech understanding for some individuals with severe sensorineural hearing loss in the 2–3 kHz range.

Keywords: Frequency compression; Hearing aid; Auditory scene analysis; Speaker recognition

1. Introduction

Luo et al. (2003) recently categorized current approaches to hearing-aid technology into three categories: signal processing to address specific impairments, attempts to increase the signal to noise ratio, and efforts to account for issues related to the user's environment. In this work, a system is proposed which addresses the first and third issues. A classifier monitors characteristics of a foreground speaker, and the classification decision is used to parametrize an existing signal enhancement algorithm which has been clinically shown to provide increased intelligibility for speakers with moderate to severe sensorineural hearing loss (Turner and Hurtig, 1999).

Auditory scene analysis is a name given to algorithms which attempt to extract information about a specific situation by the analysis of sound recordings. Typical auditory scene analysis tasks which in some cases may be augmented by visual clues include segmentation of a recording into logical units such as scenes of a film or turns in conversation, the detection of environmental noise such as office versus street noise, description of music, and content-based retrieval.

Kates (1995) was the first to consider the application of the auditory scene analysis to hearing aids. The goal of his work was to demonstrate the viability of classifying background noise into one of a number of classes such as traffic, speaker babble, printer noise, and so on. The feature set consisted of measurements representative of envelope modulation and was supplemented with linear fits of the spectrum above and below the mean. Decisions were based upon the differences in the Mahalanobis distance from centroids of the target sound classes.

More recent research has focused on the use of hidden Markov models (HMMs). Nordqvist and Leijon (2004) used ergodic discrete observation HMMs to detect speech, babble, and traffic. Observations derived from a codebook generated from cepstral coefficients and a small number of delta features were used to train models for specific noise classes. An ad hoc method used a normalized linear combination of the class dependent likelihoods as the state dependent probability in a second stage HMM whose transition weights were set empirically. The classifier would output the maximum *a posteriori* decision of the first stage HMMs and use the current state in the second stage HMM's forward decode to determine the environment in

* Corresponding author. Tel.: +1 619 594 5830; fax: +1 619 594 6746.
E-mail address: marie.roch@sdsu.edu (M.A. Roch).

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.03.002

which the first stage occurred. An example of this would be the detection of speech in traffic. Büchler et al. (2005) examined a variety of classifiers and found that ergodic HMMs offered the best performance. A number of features derived from one second intervals were explored, the best of which included tonality, width, spectral center of gravity and its fluctuation, pitch variance, and measurements of offset time.

Rather than classifying the background noise, we are interested in categorizing attributes of the foreground speaker for the purpose of enhancing his or her speech. The closest pre-existing work of which the authors are aware is a control system which was implemented for the frequency transposition system of the TransSonic FT-40 hearing aid (Parent et al., 1997). This system uses a threshold-based spectral energy magnitude peak detector. If the peak energy is located above 2.5 kHz, the system classifies the sound as a consonant and transposes the high frequency information into the lower frequency bands, effectively mixing the low and high frequency information. When the peak is below 2.5 kHz, the sound is classified as a vowel and frequency transposition is not performed.

Foreground auditory scene analysis represents a departure from other applications of auditory scene analysis to hearing aids. In addition, previous work has focused on design for current generation hearing aids. At a recent meeting of the Acoustical Society of America, Armstrong (2004) noted that Moore's law is applicable to the computational power in hearing aids. Consequently, the authors believe that it is reasonable to target research towards future generations of hearing aids and do not restrict ourselves to low dimensional feature spaces.

The remainder of this article describes an application of auditory scene analysis to hearing aids which summarizes and expands upon the authors' work previously reported at conferences (Roch et al., 2004, 2005). Differences from previously reported results are improvements in the algorithm that produce significant reductions in error rate, compensation for transducer mismatch, consideration of the use of average F0 for defining classes, and the analysis of computational complexity and error rate by phoneme class. Section 2 briefly reviews the signal enhancement algorithm. Sections 3 and 4 describe the auditory scene analysis and experiments. A discussion of the results is presented in Section 5.

2. Frequency-based compression

Turner and Hurtig (1999) designed an algorithm to perform compression in the frequency domain. The algorithm has the advantage of preserving ratios of formants without using time dilation. As suggested by Peterson and Barney's (1952) landmark study of vowels, the ratio of formants is important for perceiving speech. The patented algorithm provides linear interpolation of discrete Fourier transform bins from the original bandwidth to a reduced target bandwidth (Hurtig and Turner, 2003).

Their clinical study (Turner and Hurtig, 1999) examined 15 individuals with less than 40 dB hearing level (HL) in the 0–2 kHz range and 50–60 dB HL in the 2–3 kHz range along with a control group of three normal hearing subjects. Subjects were presented stimuli at varying compression rates (including no compression) from the UCLA nonsense syllable test (Dubno and Dirks, 1982). All stimuli were high pass amplified. Forty-five percent of the listeners showed statistically significant improvement (Student's *t*-test, 95% confidence interval (Hogg and Allen, 1978)) in comprehending compressed female speakers. For compressed male speakers, a 20% statistically significant improvement was shown.

Prior approaches to frequency modification used in hearing aids (Parent et al., 1997) had not proportionally compressed across the spectrum and the change of proportionality of the formants may counteract the advantage gained by shifting unusable frequency information into a range accessible by the hearing aid wearer. A significant drawback of Turner and Hurtig's compression method was that different classes of speech require different compression ratios. Without a control system, various classes of speech may be undercompressed or overcompressed.

3. Dynamic control and methodological framework

The need for varying compression rates has led the authors to consider a dynamic control system for the frequency-based compression. An important task is to determine classes to be recognized. In this study, the authors considered creating control classes based upon properties of specific phoneme groups, average $F0$, or biological gender.

Phoneme classes were rejected due to several difficulties. Coarticulatory effects make the analysis of classification error rate difficult. In addition, it is well known that averaging log likelihood scores produced by the same class tend to reduce the error rate. Raj and Singh (2003) formally showed that a modified Fisher ratio (F-ratio) of such a score set has a lower bound comprised of the expected F-ratio for a single score. This presents a challenge for short duration phonemes such as plosives which have average durations of 30–40 ms (Son and Santen, 1997). Using traditional speech analysis techniques with overlapping frames of length 20–25 ms results in a very small number of frames associated with some of these classes. Consequently, almost any attempt to use an averaging window would result in window lengths which are reasonably long with respect to the phone duration. The averaging window would frequently cross class boundaries, violating the assumption that all frames are from the same class.

As will be discussed in Section 4, separation of the population by average $F0$ (but not necessarily using $F0$ as part of the feature vector) produces only minor differences in both the compositions of the classes and the resulting error rate when compared to gender based classification. In either case, the rationale for separating speakers along