A Comprehensive Vowel Space for Whispered Speech

***Hamid Reza Sharifzadeh, *Ian V. McLoughlin, and †Martin J. Russell,** **Singapore, †Birmingham, United Kingdom*

Summary: Whispered speech is a relatively common form of communications, used primarily to selectively exclude or include potential listeners from hearing a spoken message. Despite the everyday nature of whispering, and its undoubted usefulness in vocal communications, whispers have received relatively little research effort to date, apart from some studies analyzing the main whispered vowels and some quite general estimations of whispered speech characteristics. In particular, a classic vowel space determination has been lacking for whispers. For voiced speech, this type of information has played an important role in the development and testing of recognition and processing theories over the past few decades and can be expected to be equally useful for whisper-mode communications and recognition systems.

This article aims to redress the shortfall by presenting a vowel formant space for whispered speech and comparing the results with corresponding phonated samples. In addition, because the study was conducted using speakers from Birmingham, the analysis extends to discuss the effect of the common British West Midlands accent in comparison with Standard English (Received Pronunciation). Thus, the article presents the analysis of formant data showing differences between normal and whispered speech while also considering an accentual effect on whispered speech.

Key Words: Whispered speech–Vowel space–British West Midlands accent–Formant analysis–Acoustic characteristics.

INTRODUCTION

Acoustic measurements of phonated vowels and diphthongs form foundational material for the speech processing and recognition fields. Wide research efforts,[1–7] mainly based on acoustic characteristics of normal vowels, show the importance of these measurements, whereas numerous studies,[8–10] in turn, have considered formant patterns in terms of vowel diagrams and the corresponding characteristics of normal vowels.

Despite the strong literature supporting normal vowels, little research effort has been spent on whispered speech relating to vowel space. Apart from the studies describing the vocal mechanism of whispers' production mostly on a glottal level,[11–14] as well as a recent study on whispered consonants,[15] the few notable studies on whispered vowels[16–18] are mainly concentrated on a few main vowels /I, ɛ, æ, ʌ, ʊ/ and conclude with general comments on vowel placement such as "higher formants in comparison with normal vowels," but accurate acoustic measurements of the precise amount of shift for each vowel/diphthong are lacking. Thus, whispered speech still lacks an acoustic vowel space determination (a classic $F2 \times F1$ plane) for researchers to refer to. Whisper vowel diagrams are useful not only for common speech processing/recognition applications but also can help those working in the biomedical engineering field of whisper-to-voice reconstruction, particularly rehabilitation of postlaryngectomized patients through restoring their normal sounding speech.[19,20]

The term "whispered speech" itself encompasses two distinct classes of speech, which we shall refer to as soft whispers and stage whispers.[21] Soft whispers (also known as quiet whispers) are produced by normally speaking people to deliberately reduce perceptibility, such as whispering into someone's ear in a theater, and are usually spoken in a relaxed manner with little effort.[11] Stage whispers, however, are used if the listener is some distance away from the speaker[21] and are actually a whispery voice, which includes partial phonation.[22] The more common soft whispers, produced without vocal folds vibration, are the focus of this study.

As mentioned, the lack of vocal folds vibration is the main physical feature and the most significant acoustic characteristic of whispered speech. It implies the absence of fundamental pitch and the harmonic relationships that are usually derived from this.[23] In a source filter model,[24] exhalation forms the source of excitation in whispered speech, with the shape of the pharynx adjusted so that vocal cords do not vibrate.[25] Exhaled air passes directly through the restricted but open larynx, causing turbulent aperiodic airflow, which forms the sound source for whispers: a rich "hushing" sound.[13]

Regarding spectral features, the spectrum of whispered speech certainly exhibits small spectral peaks at "approximately" the same frequencies as those for normally phonated speech sounds.[26] Such "formant-like" features have a much flatter power-frequency distribution[23] than normal and generally tend to be higher in frequency than the corresponding voiced speech,[27] particularly the first formant that shows the greatest difference between the two kinds of speech. Furthermore, unlike phonated vowels where the amplitude of the higher frequency formants is usually lower than the lower frequency formants, the second formants of whispered vowels are typically as intense as the first formants. Figure 1 shows this feature by contrasting the spectra of the vowel /a/ spoken in a whisper and a normal voice. These differences, mainly in first formant frequency and amplitude, are thought to be because of the alteration in the shape of the posterior areas of the vocal tract including the vocal cords, which are held rigid so as to not vibrate.[28]

The aim of this article is to propose a classic formant plane for all 11 English vowels (whispered), through analyzing the formant contours of whispered samples in a /hVd/ structure.

Accepted for publication December 6, 2010.

From the *School of Computer Engineering, Nanyang Technological University, Singapore; and the †School of Electronic, Electrical and Computer Engineering, The University of Birmingham, Edgbaston, Birmingham, United Kingdom.

Address correspondence and reprint requests to Hamid Reza Sharifzadeh, School of Computer Engineering, Parallel & Distributed Computing Centre, Block N4-B2A-03, Nanyang Technological University, Nanyang Avenue, Singapore 639798. E-mail: hami0003@ntu.edu.sg

Journal of Voice, Vol. 26, No. 2, pp. e49-e56
0892-1997/$36.00

doi:10.1016/j.jvoice.2010.12.002

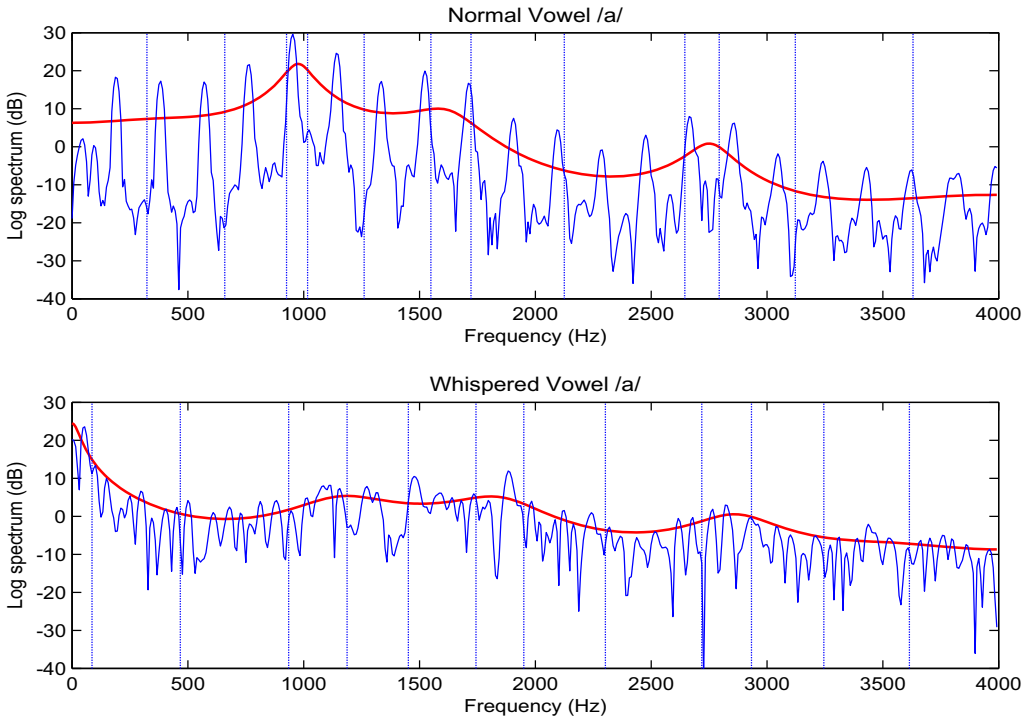

FIGURE 1. Comparison of the spectra for vowel /a/ in normally phonated speech (top) with whispered speech (bottom) for a single listener during a single sitting. The smoothed spectrum overlay shows formant peaks existing in similar locations but less pronounced for whispered speech. Furthermore, line spectral pairs (LSPs) overlaid as vertical lines typically exhibit wider spacing for the whispered speech.

The acoustic analysis, including details of the recording, speakers, equipment, and measurement methods, is described in the "Acoustic analysis" section, whereas the "Results and discussion" section outlines the results separately for men and women as well as a brief discussion on findings including the consideration of possible accentual effects in British West Midlands (WM) accent for whispers and normal speech; finally, the "Conclusion" section concludes the article.

ACOUSTIC ANALYSIS

Subjects and recordings

Speakers of this study consisted of 10 middle-aged volunteers (5 men and 5 women, 35–45 years old) born and living in Birmingham all their lives. An additional criterion of one's parents having lived in the area most of their lives was also used for the selection of volunteers.

Audio recordings were made of subjects' reading lists containing 11 vowels (/I, i, ɛ, æ, ɑ, Λ, ɒ, ə, ɔ, ʊ, u/) in a soundproof audio booth, five times with normal phonation and five times in whispered mode (total 10 times).

Subjects read from five different randomizations of a list containing the words "heed," "hid," "head," "had," "hard," "hudd," "hod," "heard," "hoard," "hood," and "who'd." Because the objective is to find out how ordinary people from Birmingham speak the vowels in the specific words, /hVd/ carrier gives actual and meaningful words in most cases, except "hudd" (which this also occurs as a family name) while this also keeps the present study aligning with previous acoustic studies on vowels[8–10] through following the same pattern.

Furthermore, having a plosive phoneme such as "d" at the final syllable makes it simple to detect vowels in between from carriers within both automatic or manual methods of extraction; particularly, because of showing a peak of energy in both whispered and spoken modes after a very short silence, "d" can be a good choice for the final syllable.

Recordings were made of five readings of the list in each whisper and normal modes (total $5 \times 11 \times 2 \times 10 = 1100$ samples). The details of the interface are described in the "Equipments and interface" subsection. If the subjects stumbled over the samples, rerecording of the samples was allowed. Speakers could repeat the sample until an accurate pronunciation was achieved.

Equipments and interface

Speech was read and recorded directly onto a laptop computer in a soundproof booth. The microphones used were an Emkay head mounted microphone and a Telex desk microphone (for near- and far-field recording, respectively). An Edirol UA-5 USB sound card interface bypassed the sound card of the laptop, removing any variation in the recordings because of different hardware. An Emkay VR3294 Battery Box provided a stable bias voltage for the microphones.

A special prompt-based recording software, developed by the University of Birmingham, was used as the recording application. Any set of prompts specified in a separate xml-formatted file with different login options can be loaded into the prompt