Hearing Research 352 (2017) 30–39

Contents lists available at ScienceDirect

Hearing Research

journal homepage: www.elsevier.com/locate/heares

Voice emotion perception and production in cochlear implant users

N.T. Jiam [a], M. Caldwell [a], M.L. Deroche [b], M. Chatterjee [c], C.J. Limb [a, *]

[a] Department of Otolaryngology-Head and Neck Surgery, University of California San Francisco, School of Medicine, San Francisco, CA, USA
[b] Centre for Research on Brain, Language and Music, McGill University Montreal, QC, Canada
[c] Auditory Prostheses and Perception Laboratory, Boys Town National Research Hospital, Omaha, NE, USA

ARTICLE INFO

Article history:
Received 5 October 2016
Received in revised form
14 December 2016
Accepted 6 January 2017
Available online 11 January 2017

Keywords:
Voice emotion
Cochlear implant
Speech prosody
Voice emotion production
Voice emotion perception

ABSTRACT

Voice emotion is a fundamental component of human social interaction and social development. Unfortunately, cochlear implant users are often forced to interface with highly degraded prosodic cues as a result of device constraints in extraction, processing, and transmission. As such, individuals with cochlear implants frequently demonstrate significant difficulty in recognizing voice emotions in comparison to their normal hearing counterparts. Cochlear implant-mediated perception and production of voice emotion is an important but relatively understudied area of research. However, a rich understanding of the voice emotion auditory processing offers opportunities to improve upon CI biomedical design and to develop training programs benefiting CI performance. In this review, we will address the issues, current literature, and future directions for improved voice emotion processing in cochlear implant users.

Contents

1. Introduction

Communicating emotion is a fundamental feature of human social interaction that transverses all cultures (Bryant and Barrett, 2008). In fact, some may argue that emotional cues formulate the very basis of human interaction and carry more valuable information that the actual words being spoken (Zajonc, 1980). There are many cues that come into play when communicating emotion, one of the most important being nonverbal cues (Skinner, 1935; Wallbott and Scherer, 1986). Among all types of nonverbal cues,

* Corresponding author. Department of Otolaryngology-Head and Neck Surgery, UCSF, 2380 Sutter St, 1st Floor, San Francisco, CA, 94115, USA.
E-mail address: charles.limb@ucsf.edu (C.J. Limb).

http://dx.doi.org/10.1016/j.heares.2017.01.006
0378-5955/

humans frequently use prosodic vocal cues (e.g., voice pitch and tempo) to elicit emotive information in their interactions (Planalp, 1996). So naturally, when prosodic vocal cues are degraded, voice emotion perception and production are often affected. Impairments in the perception and production of voice emotion usually result in serious ramifications on social interactions and social development, as in the case of infant-directed speech (Trainor and Austin, 2000), underscoring the importance of this topic at hand.

Cochlear implants (CI) are surgically implanted electrical devices that allow people with severe-to-profound hearing loss to process sound. Over the past few decades, CI development has made remarkable ground such that most CI users have adequate speech perception in quiet environments. Despite this great success, limitations remain for present day CI systems including the transmission of spectro-temporal fine structure information (e.g. pitch and harmonics) (Kong et al., 2004; Galvin et al., 2007; Kang et al., 2010; Kong et al., 2011; Xu et al., 2009). Forced to interface with highly degraded acoustic cues, CI users often demonstrate difficulty in perceiving prosodic cues. Limitations in the perception and production of prosody have adverse consequences for CI users, including the interpretation and communication of voice emotion. In this article, we will review the emerging body of work on CI-mediated perception and production of voice emotion.

2. Voice emotion general principles

2.1. Dimensions of emotions in relation to speech emotion studies

Emotions are brief and strong reactions to goal-relevant changes in the environment. Historically, there are two main approaches towards studying emotion: discrete and dimensional. A discrete approach focuses on characteristics that distinguish emotional states from one another (Ekman, 1992) whereas a dimensional approach identifies emotions based on predetermined features underlying mood and affective states (Russell, 1980). Although there are many dimensions involved in emotion, the four most commonly referred-to dimensions of subjective feeling states are activation, valence, potency, and intensity (Smith and Ellsworth, 1985). Activation refers to the perceived sense of energy ranging from low to high (e.g. somnolence to feverish excitement) (Krumhansl, 1997; Gosselin et al., 2007; Sammler et al., 2007). Orthogonally, valence relates to the intrinsic evaluation of an event, object, or situation and ranges from positive to negative (e.g. joy to displeasure) (Krumhansl, 1997; Schubert, 1999; Dalla Bella et al., 2001). Potency is a dimension used to describe the degree of powerfulness or powerlessness an individual universally identifies with a particularly emotion (Russell and Mehrabian, 1977; Osgood et al., 1957). Positive emotions almost always generate a high level of control or dominance. Thus, potency is specifically useful in differentiating between negative emotions such as fear and anger; where anger has a high potency rating and fear has a low potency rating. Last but not least, emotional intensity is used to quantify the degree of emotion being felt (e.g. very happy or only a little bit happy). The valence-arousal model approaches emotion as two separable dimensions of valence and arousal (Russell, 1980). These two dimensions are commonly used in vocal expression studies (Bachorowski, 1999; Scherer, 1986) and capture the majority of the psychophysiological components of emotion, which is why some researchers reduce emotion theory down to only two components: valence and arousal.

2.2. Speech prosody cues & voice emotion

The origin of the term ‘prosody’ can be traced back to ancient Greek where it was used to indicate the tone or accent of a syllable. Over the years, the word prosody has evolved to govern the modulation of the human voice when uttering segmental sequences of phonemes. In modern phonology, prosody refers to elements of speech relating to the properties of syllables and larger units of linguistics, such as voice pitch, duration, intensity, spectral characteristics, nonverbal vocal expressions (e.g. crying), rhythm, and tempo. Prosodic features of speech often cannot be captured by conventional segmental phonetic transcriptions or orthography. These properties of speech play an important role in communication, such as informing a listener of the speaker's intent and affect. Overall, there are few articles on the acoustic correlates of emotion. Below, we chose to highlight the most common associations found between prosodic cues and voice emotion activation and valence. The general principles mentioned in this section are not steadfast rules particularly with findings concerning valence.

Strictly speaking, pitch is the perceptual correlate of the fundamental frequency (F_0) of a sound. Although pitch is a subjective attribute and fundamental frequency an objective acoustical parameter, the two terms are often used interchangeably in the literature. For the purposes of this review, we will also use the word ‘pitch’ to refer to F_0. In speech, the fundamental frequency is derived by the rate of vocal cord vibration. The fundamental frequency range varies between speakers and depends greatly on the length and mass of the vocal cords – As a result, male speakers (85–180 Hz) generally have a lower fundamental frequency range than female speakers (160–255 Hz). Within their individual ranges, speakers have a large degree of active control over voice pitch and can choose to speak in a high or low pitch with corresponding rises and falls.

As previously mentioned, pitch is a strong acoustic cue for voice emotion in both children and adults. In fact, many school-aged children perform to the same level as adults in pitch discrimination tasks, demonstrating that fine fundamental frequency cues in voice can be available at a young age (Fig. 1) (Deroche et al., 2012). It is yet unclear whether the ability of children to recognize emotion in voice develops because this fine sensitivity to pitch is available to them very early on, or on the contrary whether voice emotion processing is one of the causes driving the auditory system to refine its sensitivity to pitch. But it is clear that the two aspects are tightly connected. For example, high activation is often associated with a high mean fundamental frequency (Breitenstein et al., 2001; Davitz, 1964; Levin and Lord, 1975; Pereira, 2000; Scherer and Oshinsky, 1977; Schröder et al., 2001) and fundamental frequency variability (Breitenstein et al., 2001; Pereira, 2000; Scherer and Oshinsky, 1977). Studies involving valence and voice pitch, on the other hand, are much less consistent. Some authors observe positive valence with low mean fundamental frequencies and high levels of fundamental frequency variability (Scherer and Oshinsky, 1977; Uldall, 1960) whereas other investigators fail to find patterns of vocal cues for valence dimension (Apple et al., 1979; Davitz, 1964; Pereira, 2000).

With most natural sounds, duration is determined by the time interval between an onset and offset. This is not only applicable for duration of sounds, but also for duration of silence between two sounds. In cases of clear and rapid changes in the sound stimuli, perception of the phonemic segmentation is more or less straightforward. With slower changes in the durations of sound segments or silence intervals, such as in glides and slurred speech, speech becomes subject to listener perception and interpretation. In general, shorter pauses are associated with high activation (Schröder et al., 2001). Longer pauses are commonly observed with negative valence (Schröder et al., 2001).

Tempo is a common acoustic cue used to convey emotion. Emotions with high levels of activation (e.g. excitement and anger) and positive valence are commonly associated with fast speech