Improvement of consonant emphasis method for bone-conduction speech intelligibility

2010025 WANG SICHENG

The recognition of bone conduction hearing is old and was known at least in antiquity. In the 1st century, Pliny the Elder, a Roman scientist, remarked on the potential of sound conduction through solid bodies. This property enables the hearing impaired to hear through the bone medium. The first bone-anchored hearing device (BAHA) became widely commercially available in the 1980s. The BAHA system uses an Osseo integrated titanium implant to propagate sound directly to the inner ear through the skull, bypassing the impedance of the skin and subcutaneous tissues. BAHA is better than conventional bone-conduction hearing aids resulting in better sound quality. Bone conduction devices are used not only in medicine but also in military and industrial applications. In this case, wearing earplugs, can effectively reduce noise damage to the outer ear and can be used for hearing protection in the military in noisy environments. Bone conduction communication can work well both in high-noise and low-noise environments. These levels are promising for the broad implementation of bone conduction communication in industrial and military applications. And also, It is expected to maintain clear communication even in high-noise environments and improve work efficiency. It is thought that bone-conducted devices can be useful for safe and secure communication. Such as the conditions of medical, firefighting, police, and other emergencies that need to safely hear the background sound and important instructions at the same time.

It has been identified that five factors contribute to bone conduction hearing: 1) sound radiated into the external ear canal, 2) middle ear ossicle inertia, 3) inertia of the cochlear fluids, 4) compression of the cochlear walls, and 5) pressure transmission from the cerebrospinal fluid. However, different from air conduction, and bone conduction has different transmission characteristics from air conduction because the part of the sound will be absorbed by body tissues.

Therefore, using a bone conduction device has a drawback. The sound quality and clarity of speech are lower when using a bone-conduction device, compared to an air-conduction one. The commonly held explanation for this phenomenon is that, under high noise conditions, the bone-conducted sound is considered to be masked by the noise heard in the air-conducted sound. Also, the bone-conducted speech's high-frequency component is attenuated due to the nature of bone-conducted transmission.

It is believed that the attenuation of high-frequency sound in bone conduction negatively impacts the intelligibility of speech transmitted through

bone conduction. Toya addressed this issue by proposing a method to enhance high-frequency sound to improve the intelligibility of speech transmitted through bone conduction in noisy environments (RT-FOE). On the other hand, Zhu investigated the results of the experiment by Toya and found that the error rate of consonants is 5 times that of vowels. Zhu focused on the time domain instead of the method proposed by Toya, which focused only on the frequency domain. Because the low power of consonants compared to vowels suggests that consonants are easily masked in noisy environments, and also easily affected by the bone conduction high-frequency attenuated characteristic. Bone conduction transmits sound through vibrations in the skin and skull, which are also transferred to the outer and middle ears. Also, the sound is transmitted from the outer ear to the middle ear by vibrations, similar to air-conducted sound. Because consonant enhancement is effective in air-conducted sounds, it is thought to be effective in the bone conduction pathway as well. In addition, an important component of the perception of consonants is the formant transition from consonant to vowel. Therefore, Zhu proposed a method to improve the intelligibility of bone-conducted speech by emphasizing the maintenance time of consonants and the formant transitions (CE). Although the performance is at its best when both methods are used simultaneously, the improvement in speech intelligibility is affected by the accuracy of consonant detection.

Through the investigation of the consonant detection results by the CE, it is found that the accuracy in consonant detection is very low. This performance of consonant detection greatly reduces the effect of consonant emphasis. By labeling the consonant segment of the database, it is found that the CE is so poor for the detection of voiced consonant segments. This study aims to further improve the accuracy of accuracy in consonant detection, by considering the characteristics of unvoiced and voiced consonants. Concretely, consonants are divided into voiced consonants and unvoiced consonants according to their vocalization patterns. Based on the characteristics of voiced consonants and unvoiced consonants, it is designed to identify voiced consonants and unvoiced consonants through power ratio. Then, the consonant detection is judged through integrated processing. Finally, also consider that the detected consonants are the perception of consonants, which is the formant transition from consonant to vowel. The formant transitions, which are important for the perception of consonants, should be emphasized together. An improved method based on CE is proposed. (CE-IMP). The difference between CE-IMP and CE in terms of consonant emphasis is that CE does not perform taper processing on the parts before the emphasis segment, while CE-IMP performs taper processing on the parts before and after the emphasis for the naturalness of speech.

This study is based on the characteristics of the power ratio to judge the unvoiced consonants and voiced consonants. Unvoiced consonants have more power at high frequencies. Based on this feature, for unvoiced consonants, use the power ratio of high-frequency power and overall power, and then compare the ratio with the threshold to judge unvoiced consonants. Voiced consonants have more power at low frequencies. Based on this feature, for voiced consonants, use the power ratio of low-frequency power and overall power, and then compare the ratio with the threshold to judge unvoiced consonants. For the thresholds related to unvoiced consonant detection and the boundary frequency of high-frequency, the best parameters are determined by the ROC curve of unvoiced consonant detection in the labeled database of unvoiced consonants and non-consonants. And also, For the thresholds related to voiced consonant detection and the boundary frequency of low-frequency, the best parameters are determined by the ROC curve of unvoiced consonant detection in the labeled database of voiced consonants and non-consonants.

To confirm the improvement effect of the proposed method on bone-conducted speech intelligibility, speech intelligibility tests were conducted in a noisy environment (55 dB, 75 dB). There are three test conditions, the first-order high-frequency emphasis compensates for the transfer characteristic of region temporalis vibration (RT-FOE), which was proposed by Fujita. consonant emphasis (CE) proposed by Zhu, and CE-IMP. According to the results of tests, in a noisy environment, CE-IMP has a significant difference in word correctness compared with other methods.