# Pitch Determination from Bone Conducted Speech

**Article** *in* IEICE Transactions on Information and Systems · January 2016
DOI: 10.1587/transinf.2015EDL8134



**2 authors:**

M. Shahidur Rahman
Shahjalal University of Science and Technology
**7** PUBLICATIONS **7** CITATIONS
SEE PROFILE

Tetsuya Shimamura
Saitama University
**211** PUBLICATIONS **887** CITATIONS
SEE PROFILE

LETTER

# Pitch Determination from Bone Conducted Speech

**M. Shahidur RAHMAN**[†a)], *Nonmember and* **Tetsuya SHIMAMURA**[††b)], *Member*

**SUMMARY** This paper explores the potential of pitch determination from bone conducted (BC) speech. Pitch determination from normal air conducted (AC) speech signal can not attain the expected level of accuracy for every voice and background conditions. In contrast, since BC speech is caused by the vibrations that have traveled through the vocal tract wall, it is robust against ambient conditions. Though an appropriate model of BC speech is not known, it has regular harmonic structure in the lower spectral region. Due to this lowpass nature, pitch determination from BC speech is not usually affected by the dominant first formant. Experiments conducted on simultaneously recorded AC and BC speech show that BC speech is more reliable for pitch estimation than AC speech. With little human work, pitch contour estimated from BC speech can also be used as pitch reference that can serve as an alternate to the pitch contour extracted from laryngograph output which is sometimes inconsistent with simultaneously recorded AC speech.

***key words:*** *bone conducted speech, air conducted speech, pitch frequency, laryngograph output, pitch reference*

## 1. Introduction

Pitch (fundamental frequency) determination of speech signals is essential in many applications of speech processing, such as speech coding, automatic speech recognition, and noise suppression. Numerous pitch determination algorithms (PDA) [1]–[5] have also been developed, none of which, however, is perfect. Further, a common shortcoming of all types of the extracted features is the sensitivity to background noise. Recently BC speech has received a lot of attention by many researchers [6]–[8], where the voice is recorded by placing a bone-conductive microphone on the talker's head. Since the bone conduction pathways (i.e. vocal tract wall and skull bone) pass only the low frequency components, BC speech is not that affected by the vocal tract resonances. Moreover, the bone-conductive microphone captures only the bone vibrations and so is thus less susceptible to background noises. Though an appropriate mathematical representation of the BC speech is still unknown, the signal of voiced BC sound is periodic, while an unvoiced signal is noiselike. Short-time autocorrelation peak, short-time energy and short-time zero-crossing rate computed from BC speech also show similarity with those of AC speech. Experimental results show that pitch determined from BC speech provides accurate estimates that could be very useful for robust speech processing applications.

Again, since no perfect PDA is available, performance evaluation of PDAs is also very important. The best way to evaluate a PDA is to use an interactive technique but this procedure needs a lot of human work. Another evaluation approach is to use a well-known method, whose performance was known *a priori* to be good, and compare the target PDA to the results of that one. To produce reference pitch contours, Hess and Indefrey [9] proposed a technique that does not derive pitch from the speech waveform, but determines it directly from the glottal vibration. They employed the output signal of Laryngograph, that monitors the vocal fold activity in the larynx. Every glottal cycle is represented by a single pulse, thus free from gross errors. Due to this benefit this method appears promising to yield a pitch reference. Unfortunately, in practice, inconsistency is observed in simultaneous recordings of speech and laryngograph signal [10]. Sometimes the speech waveform sounds voiced while the laryngograph shows no activity. Similarly, laryngeal activity does not always cause a speech waveform. This mismatch requires exclusion of the uncertain frames from evaluating algorithms. In this paper, we propose to utilize BC speech for producing pitch reference database. We studied simultaneously recorded AC and BC speech spoken by four female and four male speakers and found that completely reliable and consistent pitch reference can be produced from BC speech.

## 2. Characteristics of Bone Conducted Speech

When we speak, voice signal is transmitted via two different paths. Air conduction is the normal path of the sound that exits from mouth and transmitted through air. Bone conducted component, on the other hand, travels as vibrations through the vocal tract wall and skull bone on its way to the cochlea. A headset directly coupled with the skull of the talker can capture this bone conducted signal.

Ignoring the effect of lip radiation for simplicity, AC speech can be modeled as

$$x = u * v \tag{1}$$

where $u$ and $v$ represent glottal waveform and vocal tract impulse response, respectively. The operator $*$ stands for

Manuscript received June 11, 2015.
Manuscript revised September 13, 2015.
Manuscript publicized October 1, 2015.

[†]The author is with the Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet 3114, Bangladesh.
[††]The author is with the Graduate School of Science and Engineering, Saitama University, Saitama-shi, 338–8570 Japan.
a) E-mail: rahmanms@sust.edu
b) E-mail: shima@sie.ics.saitama-u.ac.jp

DOI: 10.1587/transinf.2015EDL8134

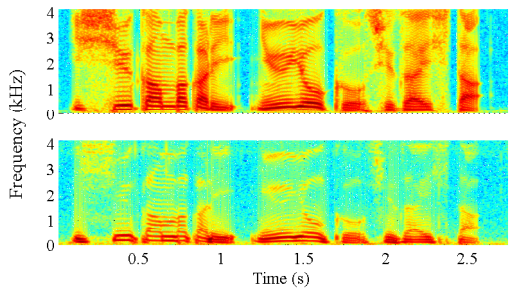


**Fig. 1** Spectrogram of simultaneously recorded AC (Top) and BC speech (Bottom).

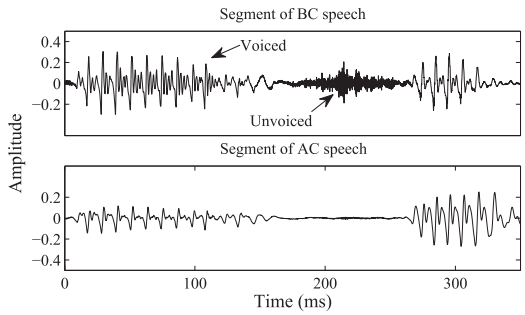


**Fig. 2** A segment of simultaneously recorded AC (Top) and BC speech (Bottom).

convolution. For voiced sound, the source is quasiperiodic puffs of the airflow through the glottis vibrating at a certain fundamental frequency. BC speech, on the other hand, can be defined as

$$y = u * v * b * k * m \tag{2}$$

where, $b$, $k$ and $m$ represent skull bone, skin and microphone impulse response, respectively. Equation (2) can be rewritten using Eq. (1) as

$$y = x * b * k * m \tag{3}$$

which indicates that compared to AC speech $x$, BC speech $y$ is influenced additionally by the bone and skin properties. Researchers sought to identify the characteristics of BC speech [7], [11]. When the sound vibrations propagate through the skull bone, these need to overcome the bone's opposition to transfer energy caused by its impedance. Due to the combined effect of bone and skin impedance, spectral characteristics (i.e. $v$) of $x$ is modified that yields sounds at lower frequencies passed on during bone conduction while sounds at higher frequencies are impeded. Spectrogram of simultaneously recorded AC and BC speech signal of a male utterance $S1$ ("isshukan bakari nyuyoku wo shuzaishita") and a selected portion of voiced and unvoiced speech signal are shown in Fig. 1 and Fig. 2, respectively. As obvious both in Fig. 1 and Fig. 2, voiced BC signal is periodic with the same periodicity as the AC counterpart and unvoiced signal is noiselike. Further, voicing alignments of both type of speech agree with each other. Though, the amplitude of the

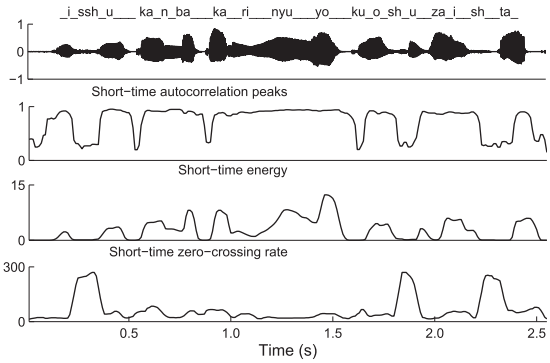


**Fig. 3** Short-time autocorrelation peaks, short-time energy and short-time zero-crossing rate obtained from AC speech.

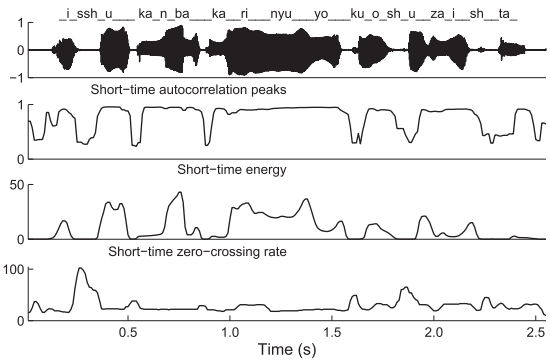


**Fig. 4** Short-time autocorrelation peaks, short-time energy and short-time zero-crossing rate obtained from BC speech.

BC speech is varied compared to AC speech depending on the frequency characteristics of AC speech [12], it does not affect the pitch properties of BC speech. The peak of the short-time autocorrelation function, short-time energy, and short-time zero-crossing rates are three simple time-domain measurements that are often utilized in voicing classification and pitch determination of speech signal. Voiced speech is characterized by relatively high energy and relatively low zero-crossing rate while unvoiced speech will have relatively high zero-crossing rate and relatively low energy. A higher value of the first peak of the autocorrelation function and the associated location provides an attractive measure of signal voicing and periodicity. These three measurements computed from the AC and BC speech signals of the utterance $S1$ are shown in Figs. 3 and 4, respectively. As in the case of AC signal in Fig. 3, the higher value of the correlation peaks and energy in Fig. 4 correspond the voiced region and higher zero-crossing rates corresponds the unvoiced region. This confirms the fact that the voicing and periodicity information of AC speech are retained in the BC speech. All the speech materials used here are recorded at 48 kHz rate which is then down sampled to 12 kHz for processing. A standard Panasonic RP-VK25 microphone is used for recording AC speech and a Temco HG-17 microphone is used for capturing BC speech where the vibration sensor is positioned on the top of the head (vertex).

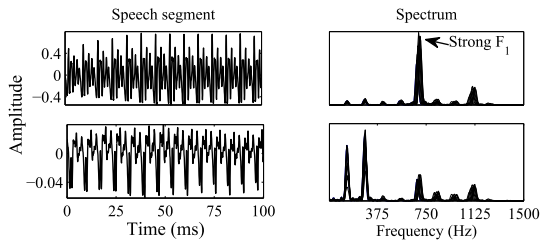


**Fig. 5** Strong $F_1$ in AC speech spectrum (Top) which is absent in case of BC speech (Bottom).

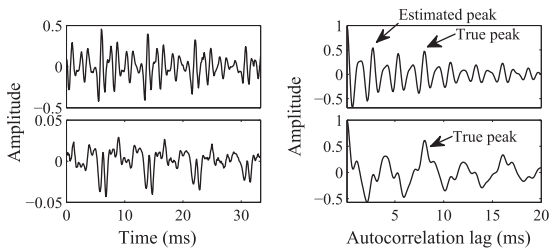


**Fig. 6** Error in pitch detection due to strong $F_1$ in case of AC speech (Top) and accurate pitch detection in case of BC speech (Bottom).

## 3. Pitch Determination from Bone Conducted Speech

In trivial condition, pitch frequency estimated from BC and AC speech are same. Average of the pitch estimates obtained from the voiced frames from AC and BC speech signals given in Fig. 2 are found to be 89.8 Hz and 90.6 Hz, respectively, which are essentially same. However, pitch estimation from AC speech is not perfect for every voice and environmental conditions. For example, if there is a strong first formant ($F_1$) present in the signal, it often causes gross errors. On the other hand, BC speech is inherently much less affected by strong $F_1$. Spectra estimated from few consecutive 40 ms frames of simultaneously recorded AC and BC signal of vowel sound /a/ at 10 ms interval are shown in Fig. 5 together with time-domain signal. /a/ is an open vowel with higher $F_1$ and as seen in the Fig. 5, the spectrum is influenced by a strong $F_1$. Autocorrelation of one of the frames are shown in Fig. 6, which is observed to introduce gross pitch error (GE) in case of AC speech. In contrast, such effect is mostly absent in case of BC speech which is thus very much suitable for pitch determination.

## 4. Pitch Reference Produced from Laryngograph Signal

The Laryngograph signal is obtained by measuring the electrical impedance at the level of the larynx by placing two electrodes on the skin on either side of the larynx. When the vocal folds apart, current flow is at a minimum and as the vocal folds snap together the current flow rises rapidly that yield a complete glottal cycle. Though the slowly varying DC is present in the output, every normal cycle is represented by a single pulse which is very suitable for pitch

**Table 1** An analysis on Keele pitch database

| Spkr. | $V(n)$ | $UV(n)$ | $C_L(n)$ | $C_S(n)$ | $\%C_F(n)$ |
|---|---|---|---|---|---|
| MA1 | 1776 | 1725 | 130 | 105 | 6.7 |
| MA2 | 1234 | 1588 | 100 | 265 | 12.9 |
| MA3 | 1416 | 1192 | 12 | 97 | 4.2 |
| MA4 | 1553 | 1582 | 39 | 196 | 7.5 |
| MA5 | 1960 | 1854 | 35 | 181 | 5.7 |
| FE1 | 1421 | 1620 | 55 | 125 | 5.9 |
| FE2 | 1813 | 1357 | 14 | 186 | 6.3 |
| FE3 | 1414 | 1443 | 25 | 168 | 6.8 |
| FE4 | 1620 | 1275 | 50 | 215 | 9.1 |
| FE5 | 1786 | 1950 | 49 | 85 | 3.6 |

determination free from GEs. In theory it looks trivial, however, in practice there are serious problems. The electrodes, for example, can loose its contact with larynx as the speakers move. This causes inconsistency between speech signal and the laryngograph output. Significant number of such uncertain frames can limit the scope of the pitch reference produced from the laryngograph signal. An analysis is presented in Table 1 using the well-known Keele pitch reference database [10]. The core database contains a simultaneous recording of speech and laryngograph output for a phonetically balanced text which was read by 10 speakers, five male (MA1∼MA5) and five female (FE1∼FE5). The second and third column indicate the number of matched-voiced ($V(n)$) and match-unvoiced ($UV(n)$) frames where the voicing classifications (voiced or unvoiced) of speech and laryngograph signal are matched. The fourth column indicates the frame numbers ($C_L(n)$) where the speech is voiced but the laryngograph trace is corrupted. The fifth column shows the frame numbers ($C_S(n)$) where laryngeal activity does not yield a speech waveform. Finally, the last column represents the percentage of total mismatched frames ($C_F(n)$) with respect to matched-voiced and -unvoiced frames. Upto 12.9% mismatch between laryngograph output and normal speech signal is observed in the table. This indicates that building a pitch reference using laryngograph signal requires reasonable human-work.

## 5. Pitch Reference Produced from Bone Conducted Speech

Bone conducted speech is induced by the sound vibrations arriving at the head through the vocal tract wall and bones of the skull. Thus BC speech can be treated as a filtered version of AC speech. When there is AC signal, there is BC signal. Voicing mismatch (as happens for the laryngograph output with AC speech) does not take place in case of BC speech. Further, compared to AC speech, BC speech is less affected by the dominant first formant and much less susceptible to background noise. Most importantly, due to its lowpass nature, the first harmonic of the BC speech spectrum can be readily emphasized to be dominant, which is actually a precondition of accurate pitch determination. Product of the original spectrum with one and two of its decimated (by 2 and 3) versions obtained from a 40 ms frame (extracted from the segment shown in Fig. 5) as described in [13] are shown

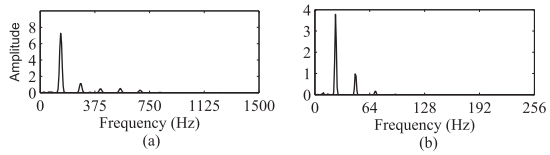


**Fig. 7** Product of the original spectrum (a) with one of its decimated (by 2) version, (b) with two of its decimated (by 2 and 3) versions obtained from a 40 ms frame.

**Table 2** Comparison of accuracy in pitch estimation using BC and AC speech

| Spkr. | $V(n)$ | AC | CC | YIN | SWIPE | EBIFF |
|---|---|---|---|---|---|---|
| MA1 | 948 | 8 / 44 | 20 / 57 | 17 / 41 | 3 / 42 | 25 / 96 |
| MA2 | 922 | 21 / 41 | 21 / 49 | 11 / 35 | 1 / 24 | 16 / 27 |
| MA3 | 546 | 13 / 51 | 18 / 47 | 18 / 48 | 2 / 35 | 19 / 41 |
| MA4 | 1017 | 14 / 38 | 8 / 45 | 8 / 17 | 2 / 8 | 18 / 22 |
| FE1 | 1042 | 15 / 36 | 23 / 37 | 14 / 20 | 2 / 15 | 29 / 24 |
| FE2 | 1162 | 14 / 42 | 19/ 44 | 24 / 39 | 3 / 42 | 31 / 41 |
| FE3 | 786 | 15 / 19 | 13/ 21 | 12 / 12 | 1 / 5 | 23 / 14 |
| FE4 | 961 | 12 / 22 | 23 / 28 | 10 / 30 | 2 / 23 | 13 / 19 |

in Figs. 7 (a) and 7 (b), respectively. As obvious in the figures, product of only two or three harmonic spectra of BC speech is adequate for the first harmonic to be most emphasized. In summary, due to the above features BC speech appears to be very much appropriate for estimating reference pitch contours. Accuracy of pitch estimation is evaluated using simultaneously recorded AC and BC speech spoken by four male and four female speakers, where every speaker utters four phonetically balanced Japanese sentences. The reference file is constructed by computing the pitch frequencies from both AC and BC speech every 10 ms using a semi-automatic technique based on visual inspection, where the initial pitch estimates are obtained using a state-of-the-art algorithm described in [4]. Pitch estimation error is calculated as the difference between the reference and estimated pitch frequency. If the estimated pitch for a frame deviates from the reference by more than 20%, we recognize the error as GE. The evaluation result is presented in Table 2 for different pitch estimation algorithms. The second column in the table shows the number of voiced frames ($V(n)$) detected in the respective utterance. From the third column on, the numbers represent the GEs obtained from BC and AC speech ($GE_{BC}(n)/GE_{AC}(n)$), respectively, using Praat's AC and CC [14], YIN [3], SWIPE [4], and EBIFF [5] algorithms. As indicated by the GEs, BC speech produces much less errors compared to AC speech for almost all the methods. Since Praat uses a different frame length from the others, pitch estimates at the boundary frames sometimes does not agree with the reference pitch. Again, results obtained using SWIPE algorithm are seemed to be little biased because the reference file is constructed based on the primary estimates obtained using this method. Though the instantaneous frequency based method [5] works by zero-frequency filtering, it is observed that the filtered signals produced by applying the algorithm on BC speech is more suitable for pitch estimation than AC speech. To summarize, when signals from both the AC and BC channels are available, pitch determination from BC speech can yield superior performance. With little human-work pitch contour extracted from BC speech may also be used as pitch reference for evaluating the performance of a PDA as currently served by the laryngograph output. Again, compared to laryngograph output, BC speech is robust to speaker movement. It does not prevent speaker from natural articulation and hearing. Voicing states (voiced/ unvoiced) are also completely consistent with AC speech.

## 6. Conclusion

Accuracy of pitch detection from BC speech has been investigated in this paper. We observed that BC speech is inherently free from the effect of dominant first formant. Due to this property, the BC speech can be readily processed to yield more accurate pitch estimates than AC speech. Moreover, BC speech is less sensitive to background conditions. Experimental results showed that with little human work the pitch contours extracted from BC speech can be used as pitch reference for evaluating the performance of a PDA. This creates an alternate to the pitch reference constructed from laryngograph output which is sometimes inconsistent with AC speech.

### References

[1] W. Hess, Pitch determination of speech signals, Springer, 1983.
[2] T. Shimamura and H. Kobayashi, "Weighted Autocorrelation for Pitch Extraction of Noisy Speech," IEEE transactions on speech and audio processing, vol.9, no.7, pp.727–730, Oct. 2001.
[3] A. De Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," The Journal of the Acoustical Society of America, vol.111, no.4, pp.1917–1930, 2002.
[4] A. Camacho and J.G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," The Journal of the Acoustical Society of America, vol.124, no.3, pp.1638–1652, 2008.
[5] B. Yegnanarayana and K.S.R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," IEEE Trans. Audio, Speech, Language Process., vol.17, no.4, pp.614–624, 2009.
[6] S. Reinfeldt, P. Östli, B. Håkansson, and S. Stenfelt, "Hearing one's own voice during phoneme vocalization-Transmission by air and bone conduction," The Journal of the Acoustical Society of America, vol.128, no.2, pp.751–762, 2010.
[7] M. McBride, P. Tran, T. Letowski, and R. Patrick, "The effect of bone conduction microphone locations on speech intelligibility and sound quality," Applied Ergonomics, vol.42, no.3, pp.495–502, 2011.
[8] E. Uchino, K. Yano, and T. Azetsu, "A self-organizing map with twin units capable of describing a nonlinear input–output relation applied to speech code vector mapping," Information Sciences, vol.177, no.21, pp.4634–4644, 2007.
[9] W. Hess and H. Indefrey, "Accurate pitch determination of speech signals by means of a laryngograph," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol.9, pp.73–76, 1984.
[10] F. Plante, G. Meyer, and W.A. Ainsworth, "A pitch extraction reference database," Proc. EUROSPEECH, pp.837–840, Madrid, 1995.
[11] S. Stenfelt and R. Goode, "Transmission properties of bone conducted sound: measurements in cadaver heads," Journal of the

Acoustical Society of America, vol.118, no.4, pp.2373–2391, 2005.
[12] M.S. Rahman and T. Shimamura, "A study on amplitude variation of bone conducted speech compared to air conducted speech," Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp.1–5, Oct. 2013.
[13] M.R. Schroeder, "Period Histogram and Product Spectrum: New Methods for Fundamental-Frequency Measurement," The Journal of the Acoustical Society of America, vol.43, no.4, pp.829–834, 1968.
[14] P. Boersma and D. Weenink, Praat: Doing Phonetics by Computer (Ver 5.1.32).