Acoustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Ambiguity in the recognition of phonetic vowels when using a bone conduction microphone

Véronique Zimpfer[a] and Karl Buck[b]

[a]ISL, 5 rue du Général Cassagnou BP 70034, 68301 Saint Louis, France
[b]French German Institut of Saint Louis (ISL), 5 rue du Général Cassagnou, 68301 Saint-Louis, France
veronique.zimpfer@isl.eu

When speaking, not only air conducted noise is generated, but also vibrations can be recorded at different places on the head using accelerometers. Bone conduction microphones are less sensitive to noise than regular acoustical microphones, they can be used in harsh environments, and they are compatible with head equipment like NBC protection devices. This paper reports the first results of a study designed to evaluate the differences in perception between speech recorded with an acoustic microphone and speech recorded using bone conduction. These differences may be the cause of bad intelligibility of communication based on bone conduction microphones, even if recorded in an undisturbed environment. We studied the recognition of ten French phonetic vowels recorded with a bone conduction microphone. A listening test has been designed to show the confusions of phonetic vowels when listening to speech picked up by air or bone conduction microphones. The tests show confusion between the vowels [i] [y] [u]. All of these vowels have the frequency of their first formant in common. The spectrograms of these vowels, which are distinctly different when recorded with an aero-acoustic microphone, become almost identical when the bone conducted speech is analyzed. Moreover, the confusions of the phonetic vowels depend on the speaker.

1 Introduction

For soldiers, it is necessary to be able to communicate even in a noisy environment. Communication systems based on bone conduction have been developed for some years. Vibrators, initially developed for hearing impaired people, are also used to allow communication when wearing ear plugs. Moreover, the production of a speech signal produces vibration in the bony structure of the head which may be recorded using accelerometers [1]. The main advantages of use of bone conduction microphones are:

- Bone conduction microphones are less sensitive to environmental noise than classical acoustic microphones (especially in the low frequency domain). The Signal Noise Ratio (SNR) is better when using the bone conduction microphone than with a simple differential microphone.
- Bone conduction microphones are compatible with head equipment like NBC masks.

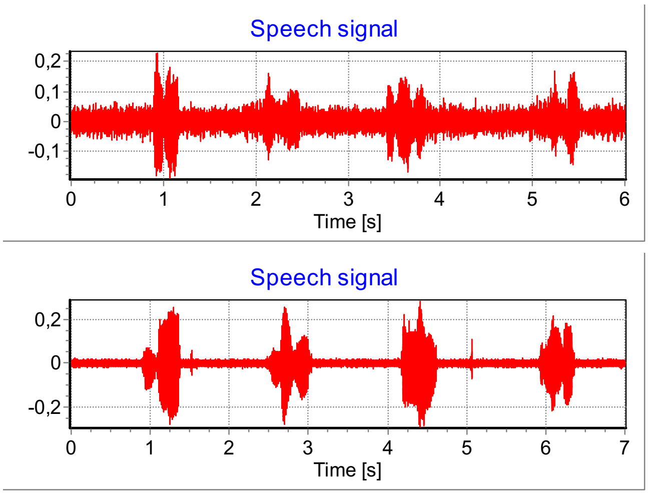


Fig.1 Speech signal of the same speaker with the standard microphone (upper plot) and with the bone conduction microphone (lower plot) in a noisy environment (pink noise at 87 dBA)

The figure 1 represents the different speech signals recorded with the standard microphone (upper curve) and with the bone conduction microphone (lower curve). The records have been realized in a reverberant room with pink noise at 87 dBA. Using the standard microphone the SNR is equal to 5 dB. On the other hand when using the bone conduction microphone the SNR is equal to 18 dB.

Due to the fact that it is less sensitive to the noise, intelligibility in noisy environments is better when using a bone conduction microphone than with a standard microphone. On the other hand in the absence of noise the intelligibility of the bone conduction microphones becomes worse. This is due to the fact that the bone conduction microphone distorts the speech signal. One of the goals of our study is to check if it is possible to use the microphones for automatic vocal recognition.

In order to evaluate the influence of the bone conduction microphone on speech recognition, a subjective and objective analysis on 10 French phonetic vowels was carried out.

2 Experimentation

Three speakers read 10 French vowels in an audiometric cabin. The speech signal was recorded simultaneously with a standard air microphone and two bone conduction microphones. The first bone conduction microphone is placed at the top of the head and the second at high right cheek. Each speaker was equipped with both bone conduction microphones and read a list of 10 vowels at a distance of 60 cm to a reference microphone (B&K ½”). The three speech signals were recorded simultaneously with a sampling rate of 24 kHz. Each speaker read a different list of the same vowels:

- List 1 (speaker 1): [a] [ã] [o] [u] [e] [i] [ɔ̃] [y] [ɛ̃] [ø],
- List 2 (speaker 2): [ɛ̃] [o] [e] [i] [ɔ̃] [y] [a] [u] [ø] [ã],
- List 3 (speaker 3): [u] [ø] [i] [e] [a] [y] [ɛ̃] [ã] [ɔ̃] [o].

3 Results of subjective test and discussion

In order to estimate the intelligibility of the vowels, a subjective test was carried out for one of the two bone conduction microphones.

		Spoken phonetic vowels									
		[a]	[ø]	[i]	[e]	[y]	[o]	[ɑ̃]	[u]	[ɛ̃]	[ɔ̃]
Recognized phonetic vowels	[a]	98%						19%		14%	4%
	[ø]		86%	2%	14%		12%	4%		2%	5%
	[i]			35%		2%			26%		
	[e]		7%		67%		21%				4%
	[y]			49%	4%	95%	7%		37%	5%	
	[o]				2%		47%				5%
	[ɑ̃]		2%					60%		5%	5%
	[u]			7%			12%		33%		2%
	[ɛ̃]	2%		2%		4%		18%	2%	74%	25%
	[ɔ̃]		2%		14%				2%		49%
	[ɛ]		4%	5%							2%

Table 1 confusion matrix of the subjective test

The test has been realized with the first bone conduction microphone placed on the high right cheek. About thirty listeners listened to the three series of vowels and noted the vowels which they recognized.

This experiment shows that the rate of recognition of the ten French vowels depends on speaker. The rate of confusion can be compared to a CVC score for which the relationship to speech intelligibility is described by Steeneken and Houtgast [2]. The rates depending on the speaker are:

- 83 % of recognition for speaker 1, rate which would be qualified “good” in a CVC (Consonant Vowel Consonant) test
- 63 % of recognition for speaker 2, rate which would be qualified “fair” in a CVC test
- < 50 % of recognition for speaker 3, rate which would be qualified “poor” in a CVC test

For speaker 1 the main confusions are:

- [i] confused with [y]
- [u] confused with [y]
- [ɑ̃] confused with [ɛ̃]

For speaker 2 the main confusions are:

- [i] confused with [y] or [u]
- [u] confused with [y]
- [e] confused with [ɔ̃]

For speaker 2 the main confusions are:

- [i] confused with [y]
- [u] confused with [y] or [i]
- [o] confused with [e] or [ø]
- [ø] confused with [e]
- Confusion between the nasal vowels

3.1 Confusion matrix

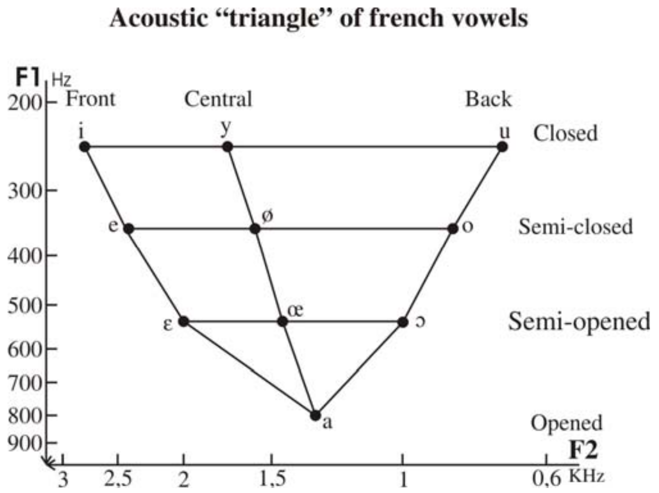
Table 1 corresponds to the matrix of confusion for all speakers. The rate of general recognition for these 10 vowels is 64%, rate which would be qualified “fair” in a CVC test.

This table shows that the three highest rates of recognition (higher than 86 %) are found for the three central vowels [a] [ø] [y] in the acoustic triangle, representation of oral vowels in the F1-F2 plan, where F1 represents the frequency of the first formant and F2 the frequency of the second formant (cf figure 2).

Confusion between the three vowels [u] [y] [i] (highlighted in green in table 1) are frequent. The common point of these three vowels is the frequency of the first formant (approximately 250 Hz) as it is displayed in the acoustic triangle of the vowels (cf figure 2). It is the same for the confusions between the vowels [o] [ø] [e] (highlighted in blue in table 1). In this case the three vowels have, at a frequency of 380 Hz, the first formant in common.

It can be concluded that the recognition of a vowel recorded with a bone conduction microphone depends on:

- The speaker: indeed for speaker 1 (series 1) few confusions appeared (rate of recognition of 83%), on the other hand for speaker 3 (series 3) many confusions were observed (rate of recognition < 50 %).
- The vowel: the vowel [i], independently of the speaker, is generally confused with [y] or [u] (rate of recognition < 40%). On the other hand, the vowel [a] is recognized up to 98 %. This high recognition rate can be explained by the fact that the vowel [a] corresponds to the tip of the vowel triangle (cf fig 2). Moreover the vowel [a] is the vowel with the highest frequency for the first formant.



,Fig.2 representation of French vowels represented in the plane of the first 2 formants (F1, F2)

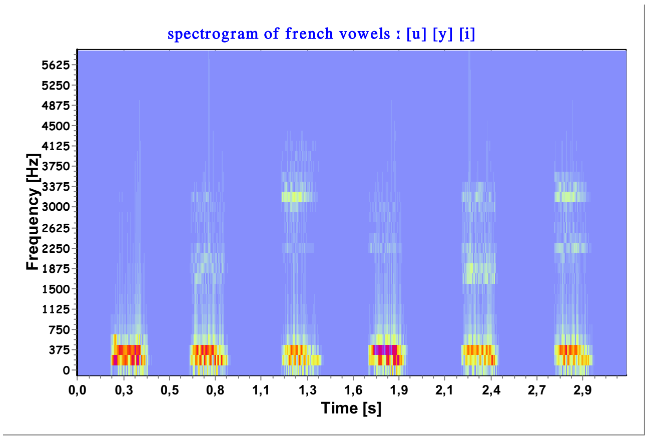
3.2 Time-frequency analyzes

For a better understanding of the results of listening tests we determined for each vowel a wide band spectrogram (Δf = 187.5 Hz). When using wide band analyzes it is possible not to take into account the fundamental of a vowel. For the groups of vowels where confusion has been detected ([u] [y] [i] and [o] [ø] [e]) the spectrograms recorded with a bone conduction microphone have been compared to those recorded with the air microphone.

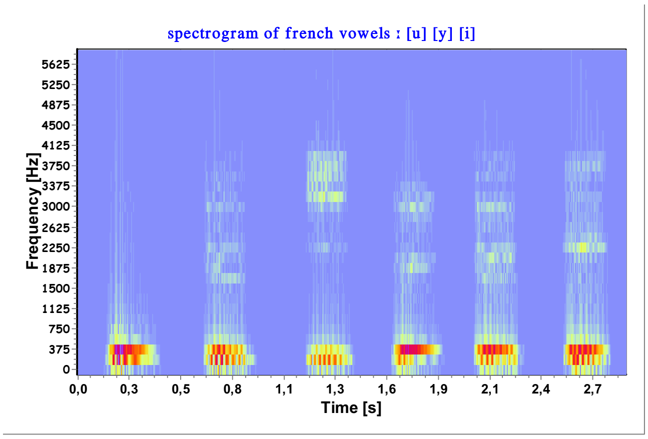
3.3 Comparison [u] [y] [i]

Figures 3 and 4 represent the spectrograms of the three vowels [u] [y] [i] for two speakers (figure 3 for speaker 1, figure 4 for speaker3). On each figure, the first three spectrograms correspond to the vowels recorded with the reference acoustic microphone. The three others are recorded with the bone conduction microphone. These figures (especially figure 4) allow to show that there is an amplification of the frequencies between 2 kHz and 2,5 kHz on the spectrograms of the speech signals recorded with the bone conduction microphone which is not visible on the signals obtained with the reference microphone.

Especially on figure 4, the spectrograms of the vowels [u] [y] [i] recorded by bone conduction microphone are almost identical and resemble the spectrogram of [y] recorded with the reference microphone. In this case, it is difficult to identify the three vowels. With the automatic speech recognition, it is only possible to recognize the central vowel [y] when one of the three vowels [u] [y] [i] is pronounced.



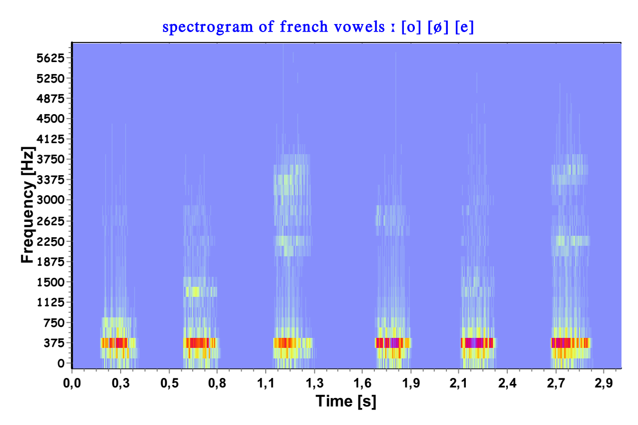
,Fig.3 Spectrogram of the vowels [u] [y] [i] pronounced by speaker 1: the first three spectrograms correspond to a measurement made with the reference microphone, the last three are recorded with a bone conduction microphone



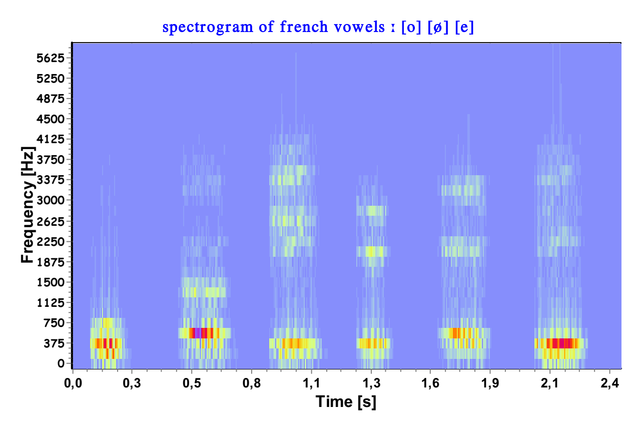
,Fig.4 Spectrogram of the vowels [u] [y] [i] pronounced by speaker 2: the first three spectrograms correspond to a measurement made with the reference microphone, the last three are recoded with a bone conduction microphone

3.4 Comparison [o] [ø] [e]

Figures 5 and 6 represent the spectrograms of the three vowels [o] [ø] [e] for two speakers. On each figure, the first three spectrograms correspond to the vowels recorded with the reference microphone and the three others to the recordings with bone conduction microphone. As for the previous case ([u] [y] [i]), these figures allow to show (figure 6) that there is an amplification of the frequencies between 2 kHz and 2,5 kHz on the spectrograms of the vowels recorded with the bone conduction microphone. Especially on figure 6, the spectrograms of the vowels [o] [ø] [e] recorded with the bone conduction microphone are almost identical, and resemble very much to the spectrogram of [e] recorded with the air microphone.



,Fig.5 Spectrogram of the vowels [o] [ø] [e] pronounced by speaker 1: the first three spectrograms correspond to a measurement made with the reference microphone, the last three are recorded with a bone conduction microphone



,Fig.6 Spectrogram of the vowels [o] [ø] [e] pronounced by speaker 3: the first three spectrograms correspond to a measurement made with the reference microphone, the last three are recorded with a bone conduction microphone

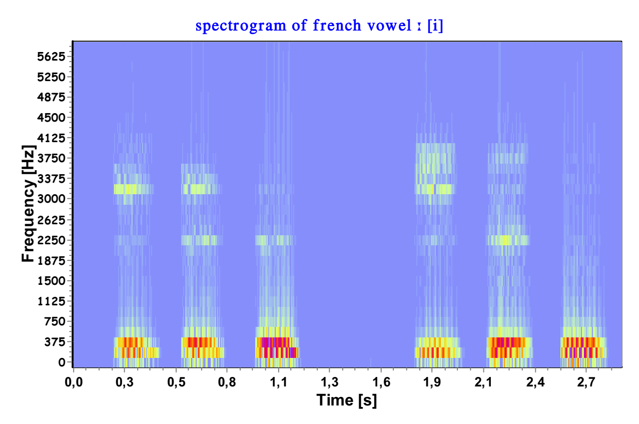
4 Comparison of two bone conduction microphones

Figure 7 represents the spectrograms of the vowel [i] for two speakers recorded with three different microphones. Figure 8 represents the spectrograms of the vowel [e] for two speakers recorded by three different microphones. On each figure, the first three spectrograms correspond to the vowel pronounced by speaker 1 and the three others to the vowel pronounced by speaker 3. For each speaker

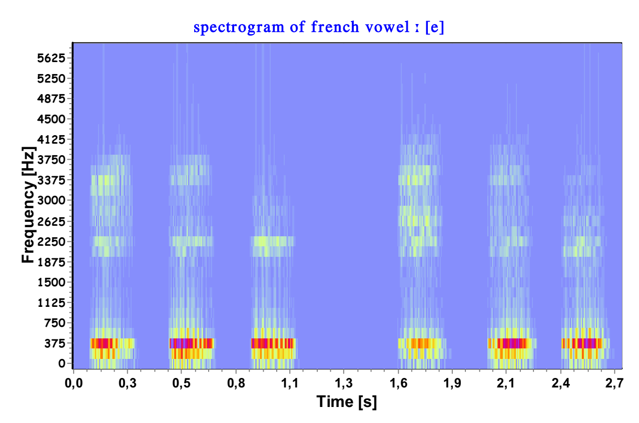
- the first spectrogram corresponds to speech signal recorded with the reference microphone
- the second spectrogram corresponds to speech signal recorded with the bone conduction microphone placed on the upper right cheek
- the third spectrogram corresponds to speech signal recorded by the bone conduction microphone placed on the forehead.

This figures show that for a vowel, the spectrogram is different in function of the used microphone. Besides, they show that spectrograms recorded with the second bone conduction microphone differ from those recorded with the first bone microphone.

The common point between these two bone conduction microphones is that the high frequency components (> 2,5 kHz) are attenuated. On the other hand the second bone conduction microphone doesn't amplify the components between 2 kHz and 2,5 kHz. But even with the second bone conduction microphone the [i] is perceived as [y]. It would be necessary to look further into the study in order to determine if the location of the bone conduction microphone on the head has an influence on intelligibility.



,Fig.7 Spectrogram of the vowel [i] pronounced by speaker 1 (first three spectrograms) and speaker 3 (last three spectrogram) recorded with the reference microphone and the two bone conduction microphones.



,Fig.8 Spectrogram of the vowel [e] pronounced by speaker 1 (first three spectrograms) and speaker 3 (last three spectrogram) recorded with the reference microphone and the two bone conduction microphones.

5 Conclusion

During the present study we have realized that recognition rate of vowels recorded with a bone conduction microphone depends on the speaker and on the vowel. It has been shown that for certain speakers the recognition rate is quite bad, whereas for others it may be satisfactory. It has been noticed that in the triangle of French vowels (figure 2) only the central vowel shows for all speakers a satisfactory recognition rate.

When analyzing the vowels with large band spectrograms it has been seen that the frequencies in the range between 2 kHz and 2.5 kHz are amplified when compared with the spectrogram calculated from signals recorded with aero-acoustic microphones. This amplification may be the major reason for the confusions between the vowels [o] [ø] [e] and between the vowels [u] [y] [i].

To answer the question about the feasibility of using a bone conduction microphone for future functions such as automatic speech recognition,

In one hand:

- Bone conduction microphones are less sensitive to environmental noise, the system will have a greater performances in strong environments.
- They are compatible with head equipments such as NBC masks.

In an other hand,

- It is certain that in the case of speaker 3, the system will have difficulties to distinguish between an [i] or an [u] and an [y].
- A speech recognition engine, being based on the recognition of the vowels, will be able to distinguish in certain cases only three families of vowels:
  - [a],
  - [ø] for the vowels [o] [ø] [e] (and even for [ɔ] [œ] [ɛ]),
  - [y] for the vowels [u] [y] [i].
- To be able to use the bone conduction microphones it is necessary to choose the vocabulary in a way that there are no ambiguities, i.e. not to choose words whose continuation of the vowels is the same one when considering that only 3 vowels ([a] [ø] [y]) are recognized.

Future investigations will be realized to understand,

- the mechanisms leading to different spectrograms of the speech signals when recorded by bone or by air microphones
- the importance of the position of the accelerometer on the head.

Acknowledgments

The acknowledgments go to SAGEM DS which contributed to the study and which provided us bone conduction microphones.

References

[1] Paolo E. Giua, "Voice transmission through vibration pickups", CNR Istituto di Acoustica, via Cassia 1216, Roma, Italy,

[2] Steeneken H.J.M., Houtgast T.,"Basics of the STI Measurement Method", in "Past, Present and Future of the Speech Transmission Index", Edited by Van Wijngarden S.J., published by TNO Human Factors, ISBN 90-76702-02-0, 2002