# Enhancing reconstruction of bone conducted speech using adaptive filtering based on Line Spectral Frequencies

Course: 02490

Stefan Gram & Jacob S. Vestergaard

Supervisor:
Jan Larsen

Technical University of Denmark

Handed-in:
June 7, 2010

**Project Team**


| | |
|---|---|
| Group: | Stefan Gram (sXXXXXX) |
| | Jacob S. Vestergaard (s062206) |


**Project Info**

| | |
|---|---|
| Course: | Machine Learning for Signal Processing |
| Course No: | 02459 |
| ECTS: | 5 Points |
| Period: | Spring Semester 2010 |
| Last edited: | June 4, 2010 |
| Hand-in: | June 7, 2010 |
| Supervisor: | Jan Larsen |

**Student signatures** (date: June 4, 2010)

*Stefan Gram*

*Jacob S. Vestergaard*

# Contents

# 1

# Introduction

## 1.1 Introduction

This project is inspired by the speech intelligibility problems with the special in ear INVSIO-HEADSETS headsets. These headsets are used with the Special Forces in the military, when a small concealed headset in the ear is a better option than a headset with a microphone boom arm. Headsets used for this application doesn't have to have a nice sounds quality, but does have to be understanderable (the speech intelligibility have to be good). This combined with the purer speech intelligibly the "in ear" headsets gives compared to normal headsets makes the foundation for this project.

## 1.2 Bone Conduction

Bone conduction generally means picking up the bone conducted sound through the body, i.e. the vibration from the scull in this case, which is much different from a normal microphone that picks up the sound acoustically from the air. The Bone Conduction Microphone (BCM) that INVISIO HEADSETS uses is a microphone in a rubber balloon. This makes the microphone sensitive to the vibration on the surface of the balloon, since this will compress the air inside the balloon. The balloon is then placed in the ear canal to pick up vibration from the jaw bone and thereby speech. The advantage of using a BCM instead of a regular microphone is the insensitivity to acoustic signal and thereby to external noise.

The advantage of using the bone conducted instead of the air conducted signal, is that noise from the surroundings doesn't get mixed with the voice of the headsets user. So when using an BCM, only the voice of the user is heard and not the external noise the user is in. But the problem is, that the speech signal picked up by the BCM sounds very different for a normal microphone. (It's a bit like your own voice sound when you put your fingers in your ear, which is called the occlusion effect in the hearing aid industry).

#### BCM Speech Intellibility compared to a external microphone

The sound of the bcm is more mumbled and nasal then a normal head worn microphone. The picked up speech does not have the correct balance between vowels and consonants either. There

**Figure 1.1:** *The INVISIO X5 headset*

seems to be a difference between oral and nasal sounds. Nasal sounds tend to be much more "nasal" the usual. This might be due to the pick up point (placed inside the ear and not in front of the mouth as a normal microphone) and the fact that it gets the vibration from the skull and not the acoustic sound. This could also explain why the sound generated at the lips, such as fricatives, is much weaker than they would be with a normal headset microphone placed at mouth reference point. In general it seams that following speech parameters is suffering the most in bcm compared to a normal microphone:

- Fricatives
- Voice/Unvoiced
- Nasal/Oral (Non-nasal)

It's seams like a very fair assumption here, that the difference between a bcm and a normal microphone is not constants, and is in fact time variant from phoneme to phoneme. In the case "Nasal/Oral" sounds for example. It's the balance between them that is off. The nasal sounds is much more "nasal"' then normal.

## 1.3 Motivation

Improving the speech intelligibility of a bone conducting microphone.

## 1.4 IPA vowels

To able to do this project withing the time frame available. We have limited this project to correcting the vowels, and evaluating the effect of correcting just those. To cover all possible

vowels as a training sets and have the training set have a good density/spread even through small it's small. IPA (International Phoneme Alphabet) was used as these should cover most vowels. The vowels can be described as the tongue position in two dimensions: The forward/backwards position of the tongue and the raised/lowed position of the tongue (See figure 1.2).

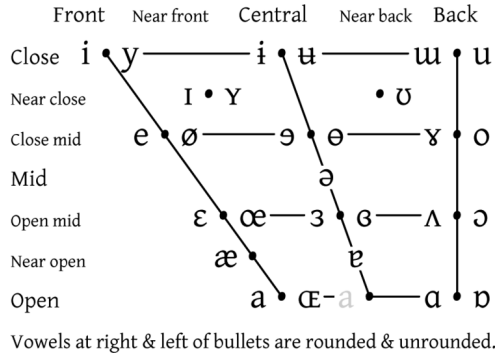


**Figure 1.2:** *IPA vowels*

FiXme Warning: Sk
ha' afsnit om hvord
lavede optagelserne
F.eks. et "Data
acquisition"' afsnit?

# 2 Analysis

## 2.1 Initial analysis

This is a brief overview of the initial part of the project, where different hypotheses were tested before moving on.

Our first hypothesis in the project was that a single stationary filter would not be adequate to reconstruct the bone conducted speech. Especially it was expected that sounds that were more nasal would need a different filtering than less nasal sounds. To investigate this, a data set of five vowels (A, E, I, O, Æ) was created, where one recording of the vowel was very nasal and the other more oral. A simple time domain plot of the recordings can be seen in Figure 2.1.

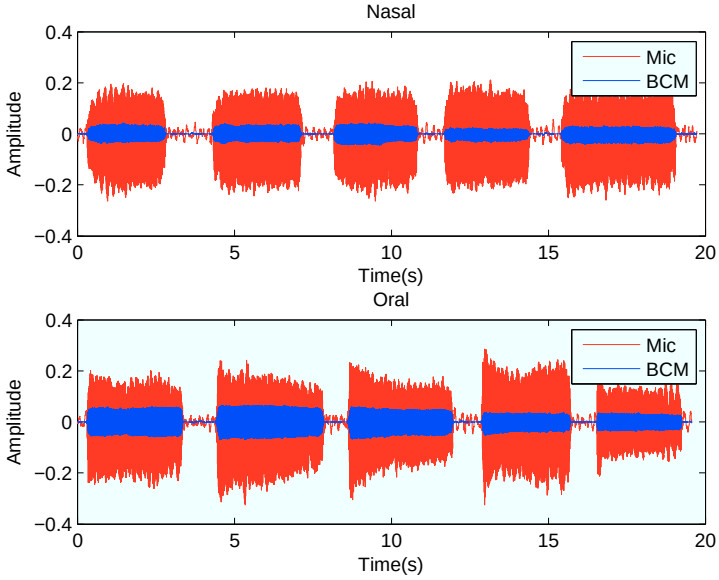


**Figure 2.1:** *Recordings of five nasal and five oral vowels (A, E, I, O, Æ)*

To affirm og discard this theory, a model to reconstruct the microphone speech from the bone conducted speech was created for each of the vowels. The model used was the Output-Error (OE) model which takes the following form

$$y(t) = \frac{B(q)}{F(q)}u(t - n_k) + e(t)$$

FiXme Warning: Polynomial order?

where the polynomials $B(q)$ and $F(q)$, in our case, are of order . We will not go further into details with this model as it was only briefly used.

In Figure 2.2 a model have been fit to a concatenation of different parts of the data set. A fit percent of the model is obtained and this is used to determine whether this type of model is adquate. For instance, *First three* means that the model is fitted on the first three nasal vowels.

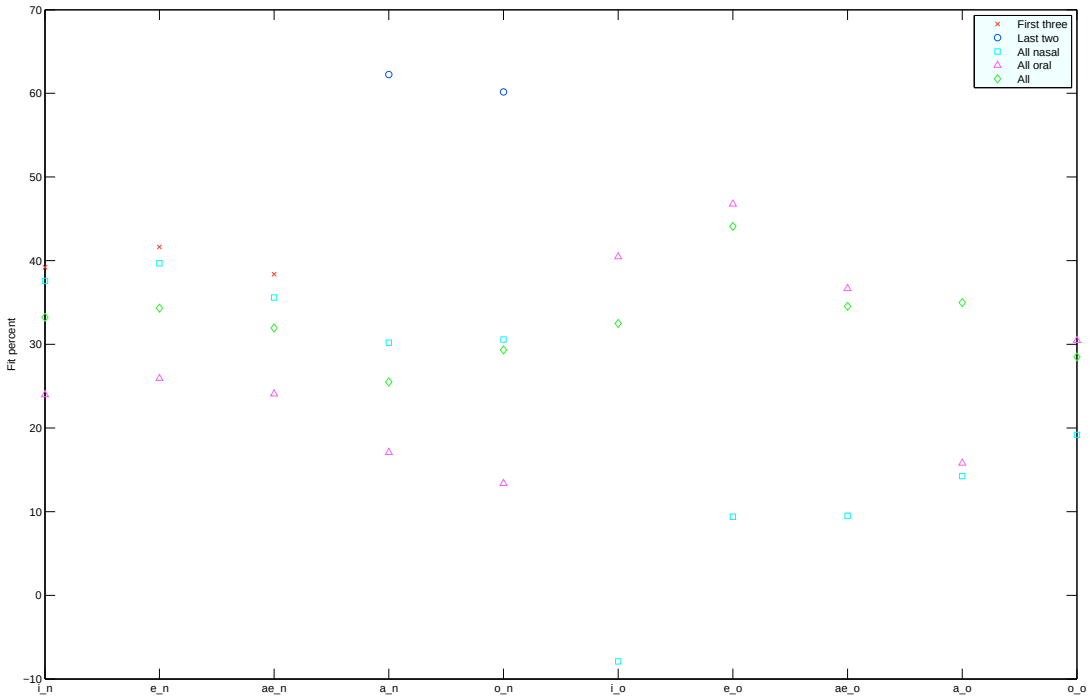


**Figure 2.2:** *Output-Error (OE) fit percent for different models*

It is seen that using either the first three vowels as training for the model and fitting to these three again provides the best model for these three vowels. This is what would be expected, as this is actually overfitting to these three sounds. The same goes for using the two last vowels. The best overall stable performance is - also as expected - seen from using all sounds as training for the model. The thing that convince us that our hypothesis that a stationary filter is not correct is the fact that a model fitted to either all nasal or all oral sounds is not capable of achieving the same fit percent when used on the other half of the set. This indicates that the dynamics in the two types of sounds are not the same.

Further modelling of single vowels separately and similar showed us that not only is a different filter for nasal and oral pronounced vowels needed, but also for different types of vowels. This increases the complexity of the reconstruction we realize that more adaptive filtering is needed.

## 2.2 Using Line Spectral Frequencies

This part of the analysis is based on (Thang et al., 2007, 2006) where it is proposed to use Line Spectral Frequencies (LSFs) to implement adaptive filtering of bone conducted speech.

**Figure 2.3:** *Example of the vocal tract $V(z)$ using LPC*

To retrieve the LSFs of a signal, a Linear Predictive Coding (LPC) model must be estimated first. Linear Predictive Coding (LPC) coefficients serve as a representation of a sound signal and is very powerful when analyzing speech. The all pole model finds the resonant frequencies of the vocal tract V(z) alone without the vocal cords E(z).

$$V(z) = \frac{Gz^{-\frac{1}{2}P}}{1 - \sum_{j=1}^{P} a_j z^{-j}}$$

This makes it easier to find e.g. the formants of a vowel.

Now, having determined the LPC coefficients of the signal, the LSFs can be retrieved. Transforming the LPC coefficient to Line Spectral Frequencies (LSF) helps making the model more robust and insensitive to the dynamics. It makes it much easier for a model to recognize the different characteristics of a phoneme. Representing the speech using LSF coefficients basically extracts the phase and orders them increasingly:

$$0 < l(1) < l(2) < \ldots < l(P) < \pi$$

For instance, raising the tone or yelling will not change the frequency of the poles, but will change how much energy they get from the vocal cord.

### Investigating capabilities of LSF

To verify that the LSF domain is in fact a domain where we can separate different vowels from each other, a data set containing ten repetitions of six vowels (A, E, I, O, U, Æ) is recorded. The LSFs are determined for these signals as described above, having chosen the order of the LPC model to $P = 12$. This results in a LSF matrix $\mathbf{X}$ of size $60 \times P$.

To visualize how well LSF space separates different vowels, we use a Principal Components Analysis (PCA) to select the three principal components that contain most of the variance and plot these against each other. This is seen in Figure 2.4 for the recordings of the BCM signal and in Figure 2.5 for the microphone signal. It is seen how similar vowels cluster in this LSF space. Especially it is seen that PC1 separates O and U from the rest, while PC2 separates these two in particular. PC3 provides good separation of I from the rest.

The fact that LSF space makes it easy to separate the different vowels from each other, we believe that this makes it a good place to do the prediction for the adaptive filtering.

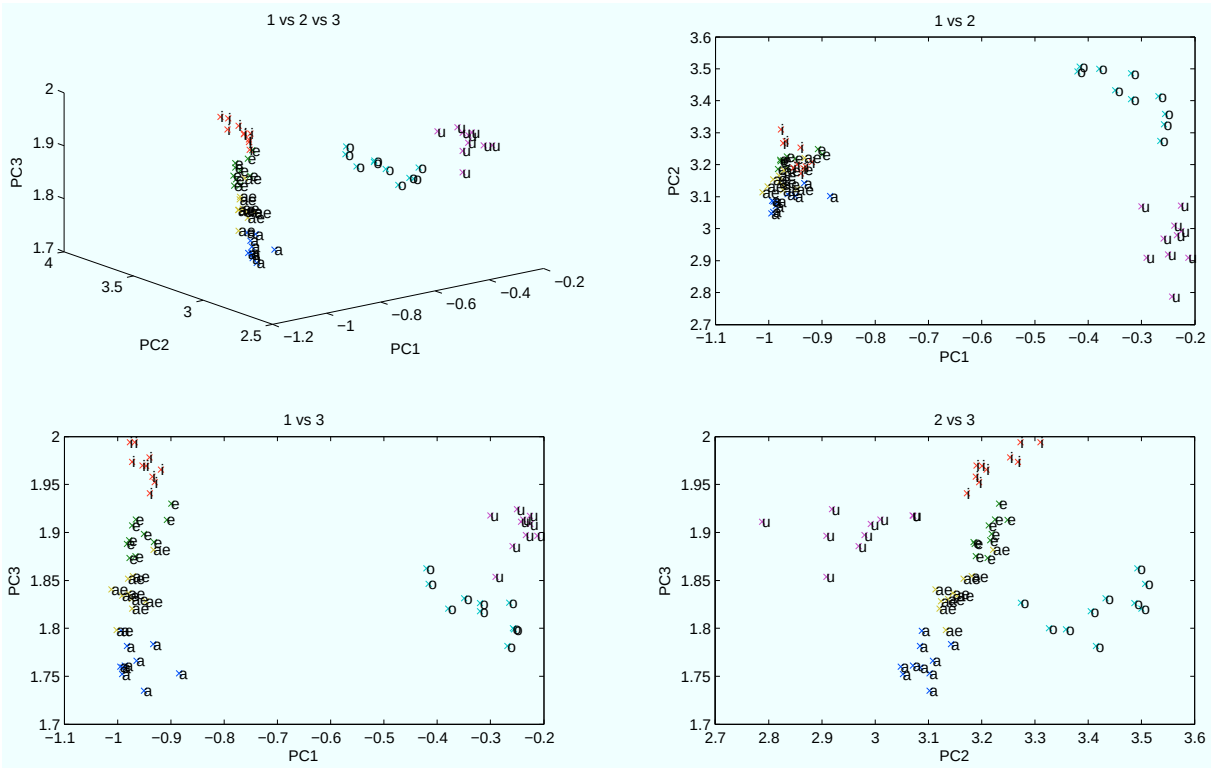


**Figure 2.4:** *First three PCs of BCM signal for ten repetitions of six vowels*

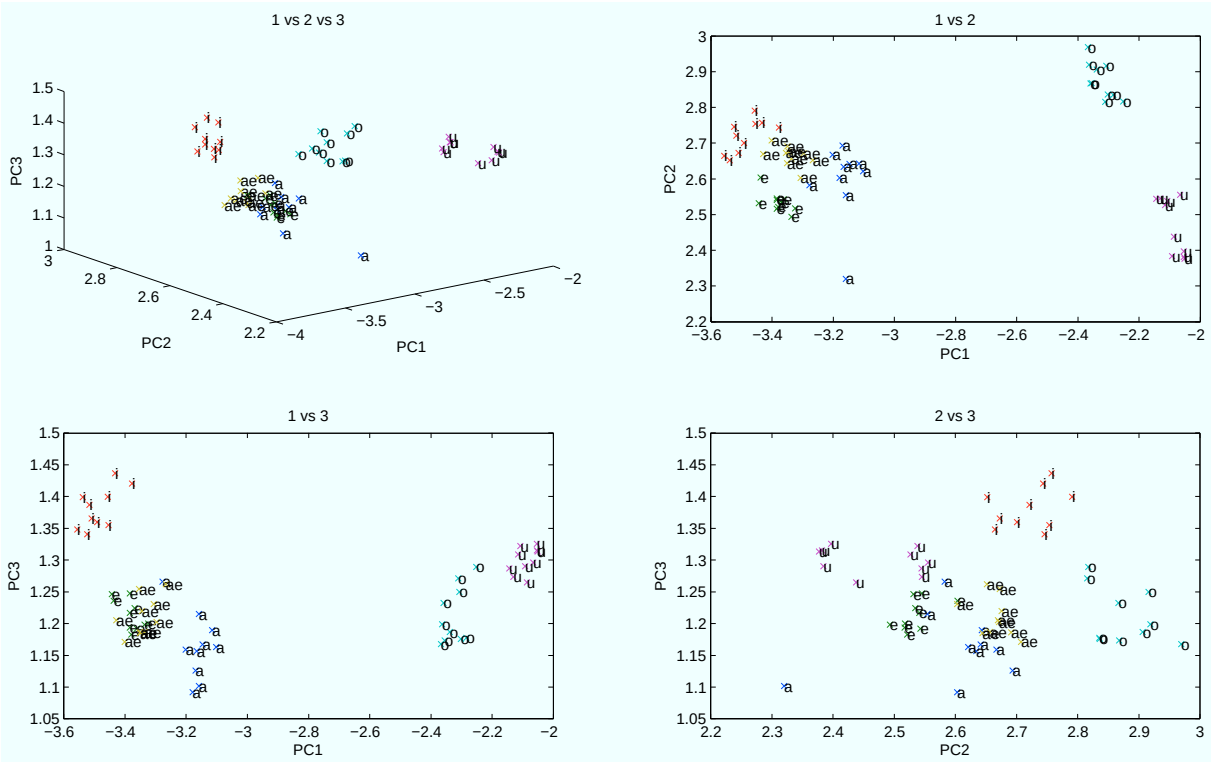


**Figure 2.5:** *First three PCs of microphone signal for ten repetitions of six vowels*

# 3

# Reconstruction model

The model we have decided to use for reconstructing the BCM signal is also based on the model proposed by (Thang et al., 2007, 2006).

## 3.1 Flow of the model

The flow of the model is outlined in Figure 3.1 and can be divided into two main parts: (1) creation of the adaptive filter and (2) the actual filtering. Here, we will first briefly describe the filtering.

**Figure 3.1:** *Flow of the reconstruction model*

## 3.2 Inverse filtering

The filter used is a standard IIR filter, where the input LPCs are used as zeros and the microphone LPCs as poles. In the ideal case we know both the true BCM signal and the corresponding microphone signal, whereby the filter is defined as

$$H^{-1}(z) = k \cdot \frac{1 - \sum_{j=1}^{P} a_{\text{bcm}}(j) z^{-1}}{1 - \sum_{j=1}^{P} a_{\text{mic}}(j) z^{-1}} \tag{3.1}$$

This would be the approach if both signals were always available. However, this is not the case, which is why we need to predict the microphone LPC coefficients.

## 3.3 Creating the adaptive filter

To create the adaptive filter we use the techniques described in the earlier sections. We estimate an LPC model for the signal and derive the LSF coefficients from these. In LSF space we do the prediction of the LSF coefficients based on a training set.

We have decided to use a very simple prediction method, namely one-nearest-neighbour. This choice has been made due to our very controlled training and test set, where we only work with vowels and phonemes. We have deliberately not chosen to use e.g. the cluster centers as the prediction, as a smooth transition between each of the vowels exists.

As decribed earlier, we use a training set of ten repetitions of six vowels, which creates a data matrix for the BCM LSF coefficients of size $60 \times 12$:

$$X_{\text{bcm}} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_{60}^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & \dots & x_{1,12} \\ \vdots & \ddots & \vdots \\ x_{60,1} & \dots & x_{60,12} \end{bmatrix}$$

We also work with a training set of the IPA as described in Section 1.4, where each of these is only repated once. This data matrix has a size of $30 \times 12$ as there is 30 different phonemes. The method described is the same - only the training set is swapped in this case.

When doing the prediction the chosen class is the nearest vowel in LSF space:

$$\text{class}(\hat{\mathbf{x}}) = \underset{i \in [1,60]}{\arg\min} \{\|\hat{\mathbf{x}} - \mathbf{x}_i\|_2\}$$

As the LSF coefficients for the corresponding microphone speech are also saved in the training set, we can now select the predicted LSF coefficients, convert back to LPC coefficients and use them to create the inverse filter from (3.1).

Finally, the input BCM signal is filtered using this adaptive filter.

## 3.4 Applying to continous speech

As part of the process we have also applied our model to continous speech. However, we have still used the vowels (or IPA) as training set. This is still from the hypothesis that vowels and phonemes are major contributors to the degradation of the BCM signal.

When applying this to continous speech, a few issues arise and especially window length is of importance. The window needs to be large enough to reduce the fluctuations and modulations arising from the filtering, but still small enough to capture the essence of the vowels. We have - after a lot of testing back and forth - used a window length of 100 ms.

To process the continous speech we divide the speech into windows of this length and create and apply an adaptive filter to these windows separately. To patch them together we use overlapping raised cosine distributions

$$f(x) = \frac{1}{2} - \frac{1}{2} \cdot \cos\left(x \cdot \frac{2\pi}{N}\right)$$

which as seen in Figure 3.2 sum to one when using 50% overlap.

**Figure 3.2:** *Raised cosine distributions with 50% overlap*

# 4

# Evaluating the model

To evaluate the model, both continous speech and simple vowels were reconstructed by using the IPA data as training set. To evaluate the quality of the reconstruction, a few objective measures were implemented as proposed in Thang et al. (2007).

## 4.1 Objective measures

The measure we have chosen to implement, each evaluate different parts of the reconstruction process. We apply the following measures by calculating them between the microphone signal and the BCM signal, before and after the reconstruction, and subtracting these two to achieve a measure of improvement.

The log-spectrum distortion (LSD) measures the numerical similarity of two frequency responses. The LP coefficient distance (LCD) measures, as the name indicates, the distance between two LP models' coefficients, which in our model is a measure of how good our training set and prediction model are. The final measure is the mel-frequency cepstral coefficient distance (MCD). This measure has similarities to the well-known cepstrums, only the MFCCs are using the mel-scale, which is logarithmic. Often MFCCs are used as features in genre classification and speech recognition, but in this case we use it as a quality measurement. We use the implementation by (Omogbenigun, 2009).

The three measures are defined as:

$$\mathrm{LSD} = \sqrt{\frac{1}{W}\sum_{\omega}^{W}\left[20\log_{10}\left(\frac{|S(\omega)|}{|\hat{S}(\omega)|}\right)\right]^2} \tag{4.1}$$

$$\mathrm{LCD} = \sqrt{\frac{1}{P}\sum_{i=1}^{P}\left(a_{\mathrm{mic}}(i) - a_{\mathrm{bcm}}(i)\right)^2} \tag{4.2}$$

$$\mathrm{MCD} = \sum_{i=1}^{1} 2\left(c_{\mathrm{mic},i} - c_{\mathrm{bcm},i}\right)^2 \tag{4.3}$$

The results from calculating the difference between the measures before and after reconstruction of the BCM signal can be seen in Table **??**. A single of the recordings of each six vowels is reconstructed.

| Training set | | A | E | I | O | U | Æ | Speech |
|---|---|---|---|---|---|---|---|---|
| Vowels | LSD | 9.1 | -10.9 | 23.5 | 18.3 | 2.0 | 1.9 | 5.25 |
| | LCD | 63.9 | 88.3 | 90.1 | 89.6 | 71.8 | 94.1 | 66.1 |
| | MCD | 75.4 | 80.8 | 92.7 | 77.6 | 90.0 | 90.6 | 19.5 |
| IPA | LSD | -1.2 | -25.8 | 2.7 | 13.6 | 18.7 | 7.6 | 7.75 |
| | LCD | 12.3 | 61.4 | 47.4 | 53.7 | -17.1 | 80.8 | 58.0 |
| | MCD | 22.9 | 52.4 | 42.3 | 60.0 | -9.5 | 74.3 | 33.3 |

**Table 4.1:** *Objective quality measurements improvement in percent*

It is seen, that when using the data set with ten repetions of six vowels, the improvement for the same vowels is larger than when using the IPA data set. This is as expected as the vowels data set is a good representation for reconstructing exactly these vowels. The interesting conclusion that can be drawn from these numbers is, that the IPA data set can in fact be used to reconstruct continous speech as good - or even better - than the vowel data set. This is important because this data set is half the size of the vowels set. This indicates that the IPA is a good representation of a training set in LSF space, as the same numerical performance can be achieved with this set.

# Bibliography

Olutope Omogbenigun. Speechcore. http://www.mathworks.com/matlabcentral/fileexchange/19298-speechcore, 2009.

Tat Vu Thang, Masashi Unoki, and Masato Akagi. A study on an lp-based model for restoring bone-conducted speech. 10(6), 2006.

Tat Vu Thang, Masashi Unoki, and Masato Akagi. Lp-based method of blind restoration to improve intelligibility of bone-conducted speech. Technical report, School of Information Science, Japan Advanced Institute of Science and Technology, 2007.