ABSTRACT
A new approach to speech normalisation is presented. A method that finds the optimal coefficient for linear slope of the warping function is described. The affine normalisation functions are suggested. Their coefficients depend on expected values of frequency when speech spectra are used as a density of probabilities. The method was developed for computer games to lower costs of recording dialogues and to make them more attractive for players.

KEY WORDS
speech normalisation, affine warping

1 Introduction
The variations in acoustic speech signals from different speakers are caused by the different sizes of a vocal tract, a gender, different accents, a dialect, a speaking rate and a style influenced by speaker’s personality and current emotional state. Vocal Tract Normalisation (VTN) is a procedure which is typically applied in speaker independent Automatic Speech Recognition (ASR). The main idea is to eliminate variation of a speech signal caused by individual features. This is achieved by warping the frequency variable of the speech spectrum. The idea of VTN has been considered by a number of authors for fifteen years [1], [2], [3], [4], [5], [7], [8].

Differences in vocal tract cause differences in spectra, even when the speakers generate a sound of the same phoneme. The warping of the frequency axis is a typical method to modify spectra so that the distance between the spectra of the same phonemes are smaller. The distance might have different meaning and yield different criteria on the choice of the warping function. The average distance in frequency domain is a good example. We studied the use of expected value of frequency for VTN to compare our method with conventional techniques.

VTN clearly improves recognition accuracy. In our case, the purpose of applying the presented below methods is different. Dialogue recordings are an important cost in a budget of computer games production. They require not only a proper studio and recording hardware to guarantee high quality speech effects. Native speakers have to provide their voices, which is a serious challenge in multilanguge production process. For example, Polish companies quite often produce games for American market. In this case, they have to order recordings from US. Supervision of such process is obviously an organisational and financial challenge. What is more, recordings for computer games are now frequently made by professional actors, often famous ones. Our aim is to reduce these costs by providing tools for speech modulation which would allow reusing recordings with various pitch tones to enable the perception as different speakers.

The aim of experiments presented below, was to modify the speech recordings and to obtain the basis of normalised speech utterances. Appropriately processed speech samples provide opportunity to make an illusion that they all have separate recordings (just by changing some parameters) making crowds in games more realistic. Such techniques can be especially useful for acoustics effects of unimportant non-player characters where recordings are made only once and represent the general population of some area.

2 Speaker normalisation technique
The important issue when examining normalisation technique is the choice of warping function. The linear warping function is very popular [6] because of its simplicity. Other functions are curves based on human perception properties, the bilinear transform or transforms based on the mel scale [6], along with speech production models [9]. Usually authors have presented speaker normalisation models using both linear and nonlinear warping functions. A linear warping function seems to capture most of the information necessary to achieve speaker normalisation. There are different conclusions for the case of speech in computer games. Some not natural speech features are frequently used to underline the specific human or machine characteristics.

The selection of warping function is usually based on maximisation of likelihood, and sometimes directly on speaker specific parameters. In the next section, we present a new normalisation technique. At the beginning, the warping function for a given speaker maximises the likelihood of the expected frequency for spectrum of a tested and a standard speaker. Next, the main part of algorithm uses
an affine warping function which depends on a coefficient
computed for the linear function.

In this work we examine the feasibility of speaker nor-
malisation by mapping directly between analysed speaker’s
and standard speaker’s amplitude spectra. A warping func-
tion in the context of speaker normalisation is a function

\[ \tilde{I} = \phi(I) \]  

(1)
mapping speech spectrum \( \tilde{s}(f) \) into spectrum \( \tilde{\phi}(\phi(f)) \).

The effect of applying continuous and strictly mono-
tonic function (1) will be an expansion for these frequen-
ties where \( d\phi/df > 1 \) or compression for \( f \) such that \( d\phi/df < 1 \).

Let us assume that we have some recorded speech
samples \( s_1(t), s_2(t), \ldots, s_I(t) \). They should be long enough
(lasting at least 20 seconds) to correctly represent the fre-
quency properties of analysed speakers. Let us assign the
complex spectra of these signals by \( \tilde{s}_1(f), \tilde{s}_2(f), \ldots, \tilde{s}_I(f) \).

The probability density functions

\[ p_i(f) = \frac{|\tilde{s}_i(f)|^2}{\int |\tilde{s}_i(f)|^2 df} \]  

(2)
can be computed for all of \( i = 1, 2, \ldots, I \) samples, where
frequency 8 [kHz] corresponds to the Nyquist frequency.
Two examples are presented in (see Fig.1).

Each speaker can be characterized by the frequency
expected values

\[ e_i = \int_0^8 p_i(f) df \]  

(3)
The standard deviation

\[ \sigma_i = \sqrt{\int_0^8 (f - e_i)^2 p_i(f) df} \]  

(4)
can be used as the additional characteristic parameter for
\( i \)-th speech sample.

In the next step, such speech sample \( \xi \) should be
found that \( e_\xi \) is a median value among all expected fre-
quencies \( e_i \).

The linear VTN function has form (see Fig.2)

\[ \tilde{f} = a_i f \]  

(5)
and coefficients \( a_i \) should satisfy equation

\[ \int_0^{f_{\max}} [p_i(a_i f) - p_\xi(f)] f df = 0 \]  

(6)
where \( f_{\max} = 8 \) [kHz] if \( e_\xi > e_i \) (i.e. \( a_i < 1 \)) or \( f_{\max} =
8/a_i \) [kHz] if \( e_\xi < e_i \) (i.e. \( a_i > 1 \)). Some results of the
speech analysis are presented in Tab.1 and Tab.2.

For all \( i \)-th samples, parameters \( a_i \) can be used to
modify their spectra to obtain \( \tilde{s}_1(a_i f), \tilde{s}_2(a_i f), \ldots, \tilde{s}_I(a_i f) \),
where \( a_i f \leq 8 \) [kHz] for all \( f \) and for all \( i = 1, \ldots, I \).
Finally we obtain the unified speech samples
\( s_1(t/a_1)/a_1, \ldots, s_I(t/a_I)/a_I \).

Instead of slope \( a_i \) it is possible to find a size of fre-
quency shift

\[ \nu_i = e_i - e_\xi \]  

(7)
and in that way to obtain the simple affine VTN function

\[ \tilde{f} = f - \nu_i \]  

(8)
Such normalisation brings usually considerable distortions
and decreases the speech quality. Although changes for
the low frequencies are acceptable, the shift volume for the
high frequencies is usually too small.

The affine normalisation function

\[ \tilde{f} = a f + b \]  

(9)
should depend on the slope coefficient \( a_i \). The next rational
assumption must be connected with coefficient \( b \). It seems
reasonable to choose such (9) that both linear (5) and affine
(9) VTN functions have the same value \( \tilde{f} \) for some chosen
frequency \( f_c \). It seems that \( f_c \) should be equal to the
greatest sensitivity of human hearing system or the speaker
frequency expected value (3).

It is a property of affine normalisation function (9)
that \( \tilde{f}(0) \neq 0 \). It makes that the lowest frequencies
must be treated in a specific way. We can assume arbitrary the
boundary values of these specific low frequency bands, i.e.
\( f_0 \) or \( \tilde{f}_0 \) (see Fig.3 and Fig.4) as coefficients for a speech
normalisation (9). Additional inequality constraints

\[ f_0 \leq f_{\max} = \frac{8 f_c (a_i - 1)}{a_i f_c - 8} \]  

(10)
and

\[ \tilde{f}_0 \leq f_{\max} = \frac{8 f_c (a_i - 1)}{8 - f_c} \]  

(11)
should hold to limit distortions caused by the speech nor-
malisation.

To summarise, for each \( i \)-th speaker, the normalisa-
tion function (9) can be uniquely determined by the coeffi-
cient \( a_i \) computed from (6), arbitrary assumed \( f_c \) and \( f_0 \) or
\( \tilde{f}_0 \). Details are presented below.

For the case \( a_i < 1 \), the normalisation function (9)
have coefficients \( a = a_i f_c/(f_c - f_0) \) and \( b = a_i f_0 f_c/(f_0 - f_c) < 0 \)
and (9) gives

\[ \tilde{f} = \frac{a_i f_c}{f_c - f_0} (f - f_0) \]  

(12)
for \( f_0 \leq f \leq 8 \) [kHz]. Inequality constrain \( f \geq f_0 \)
causes (see Fig.3) that the lowest frequencies, below \( f_0 \)
[kHz], are not transfered to the spectrum of the normalised
speech signal. The next important property of function
(12) is increasing of the frequency band of normalised
speech signal from $8a_i$ [kHz] for VTN defined by (5) to $f(8) = a_i f_c (8 - f_0)/(f_c - f_0)$ [kHz] for (12).

The second case we obtain when a solution of (6) satisfies inequality $a_i > 1$. The normalisation function (9) have then coefficients $a = a_i - f_0/f_c$ and $b = f_0$ and the final form of VTN is

$$ \hat{f} = (a_i - f_0/f_c) f + f_0 $$

(13)

for $0 \leq f \leq (8 - f_0)/(a_i - f_0/f_c)$. It means that frequencies above $(8 - f_0)/(a_i - f_0/f_c)$ [kHz] (see Fig.4) are not transferred to the normalised speech signal. It is also worth to notice that the spectrum of normalised speech does not contain frequencies below $f_0$ [kHz].

3 Experiments

At the beginning, the recorded speech samples were annotated to prepare the speech corpus for future experiments. The speech was sampled at 16 [kHz], providing signals whose spectra range up to 8 [kHz]. Speech samples from different speakers were classified according to the main frequency of their spectra (see Tab.1 and Tab.2). All speech samples were divided into two groups according to the gender of speakers.

The inverse affine mapping

$$ f = \phi^{-1}(\hat{f}) = \begin{cases} \frac{f(\hat{f} - f_0)}{a_i f_c} + f_0 & \text{for } a_i < 1 \\ \frac{(\hat{f} - f_0)f_c}{a_i f_c - f_0} & \text{for } a_i > 1 \end{cases} $$

changes the discrete frequency domain $F = \{0, \Delta f, 2\Delta f, \ldots, (N-1)\Delta f\}$ for which the spectra were computed, and $\Delta f$ is an original sampling density in the frequency domain. Mapping (14) enables us to find $\hat{s}(f)$ for $0 \leq f \leq 8$, except of the frequency values $f$ which have no equivalent $\hat{f}$ in the domain from 0 [kHz] to 8 [kHz].

For the presented experiments $\Delta f = 0.05$ [Hz] because all speech samples had the same length, equal to 20 [s]. VTN defined by (12) gives the time domain sampling frequency $\Delta t_1 = (1 - f_0/f_c)/(16a_i) > 1/16$ [ms] and VTN defined by (13) gives $\Delta t_2 = 16^{-1}(a_i - f_0/f_c)^{-1} < 1/16$ [ms].

Table 1. Data of male speech samples: expected values (3), standard deviation (4) and parameter $a_i$ which satisfies (6)

<table>
<thead>
<tr>
<th>$i$</th>
<th>$\epsilon_i$</th>
<th>$\sigma_i$</th>
<th>$a_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>274.07</td>
<td>325.71</td>
<td>1.203</td>
</tr>
<tr>
<td>2</td>
<td>288.52</td>
<td>373.06</td>
<td>1.171</td>
</tr>
<tr>
<td>3</td>
<td>351.48</td>
<td>454.55</td>
<td>1.061</td>
</tr>
<tr>
<td>4</td>
<td>389.94</td>
<td>391.68</td>
<td>1.007</td>
</tr>
<tr>
<td>5</td>
<td>395.64</td>
<td>351.81</td>
<td>1.000</td>
</tr>
<tr>
<td>6</td>
<td>408.78</td>
<td>536.91</td>
<td>0.984</td>
</tr>
<tr>
<td>7</td>
<td>445.42</td>
<td>590.02</td>
<td>0.942</td>
</tr>
<tr>
<td>8</td>
<td>464.99</td>
<td>723.38</td>
<td>0.922</td>
</tr>
<tr>
<td>9</td>
<td>467.58</td>
<td>500.27</td>
<td>0.920</td>
</tr>
</tbody>
</table>

4 Evaluations

The experiments showed that $f_c$ should be equal to $\epsilon_i$ frequency for each speaker but a bit different values are acceptable as well. All these cases preserve the similarities
of VTN defined by (5) with normalisation (12) for the case $a_i < 1$ and with normalisation (13) for the case $a_i > 1$. These similarities (see Fig.3 and Fig.4) occur for the frequency bands where speech signals have the main part of their energy (see Fig.1). The greatest differences between standard speech and speech normalised by (12) and (13) are allowed for the lowest and highest frequencies where the speech energy is much lower. The warp parameter $a_i$ varies from 0.9 to 1.2 for each gender.

The main purpose of our method is to make games more attractive. The sound effects as well as the speech signals should have the characteristic features. It causes that the objective quality measure is difficult to choose. The only way we could approach measuring of quality is by human subjective evaluation of the recordings changed by the VTN method described above.

The quality of speech samples obtained by our VTN methods were verified during a single session by 76 students. For seven speech samples, separately for each gender, students chose one of the three possible marks: 2 - perfect quality of speech, 1- some distortions are noticeable, 0 - bad quality of speech. The results are presented in Tab.3 for the male speech samples and in Tab.4 for the female speech samples. Recordings with affine frequency warping were computed for parameters $f_0$ and $f_0'$ equal to half of their maximal values according to (10) and (11).

### 5 Conclusion

Our approach results are a useful tool to make computer games dialogues much more changeable and attractive.

![Figure 3. Affine warping function (12) (bold line) for the compression case compared with linear warping function (5)](image1)

![Figure 4. Affine warping function (13) (bold line) for the expansion case compared with linear warping function (5)](image2)

<table>
<thead>
<tr>
<th>$i$</th>
<th>$c_i$</th>
<th>$a_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>334.36</td>
<td>1.133</td>
</tr>
<tr>
<td>2</td>
<td>413.10</td>
<td>1.019</td>
</tr>
<tr>
<td>3</td>
<td>416.10</td>
<td>1.016</td>
</tr>
<tr>
<td>4</td>
<td>417.46</td>
<td>1.014</td>
</tr>
<tr>
<td>5</td>
<td>429.16</td>
<td>1.000</td>
</tr>
<tr>
<td>6</td>
<td>433.32</td>
<td>0.995</td>
</tr>
<tr>
<td>7</td>
<td>436.70</td>
<td>0.991</td>
</tr>
<tr>
<td>8</td>
<td>445.48</td>
<td>0.982</td>
</tr>
<tr>
<td>9</td>
<td>505.88</td>
<td>0.922</td>
</tr>
</tbody>
</table>

Table 2. Data for female speech samples: expected values (3), standard deviation (4) and parameter $a_i$ which satisfies (6)

Obviously, the most of female speakers generate higher frequency speech spectrum than majority of male speakers. These differences between adult speakers exhibits average frequencies of around 9% higher than those of male speakers. The average change of the parameters $a_i$ for speech samples varies between 3% and 7.5%. The maximum difference between warping parameters $a_i$ for both genders is about 20%.

According to evaluation data collected in Tab.3 and Tab.4, the presented algorithms of speech normalisation do not cause significant distortions and loss of quality in the recordings. Moreover, it may be noticed that the observed decrease in speech quality for higher values of $a_i$ parameter is larger in the linear then in the affine VTN, until $f_0$ frequency is lower than 10 [Hz]. For larger values of $f_0$ parameter, speech quality drops dramatically.
Table 3. Average values of human subjective evaluations for male speech samples. Index i of speech samples is the same as in Tab.1.

<table>
<thead>
<tr>
<th>i</th>
<th>( f_{i}^{\text{max}} )</th>
<th>Average students’ marks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original speech</td>
<td>Linear (5) transformation</td>
</tr>
<tr>
<td>3</td>
<td>22.45</td>
<td>1.49</td>
</tr>
<tr>
<td>4</td>
<td>2.99</td>
<td>1.68</td>
</tr>
<tr>
<td>5</td>
<td>—</td>
<td>1.43</td>
</tr>
<tr>
<td>6</td>
<td>6.99</td>
<td>1.73</td>
</tr>
<tr>
<td>7</td>
<td>27.14</td>
<td>1.48</td>
</tr>
<tr>
<td>8</td>
<td>38.09</td>
<td>1.47</td>
</tr>
<tr>
<td>9</td>
<td>39.53</td>
<td>1.65</td>
</tr>
</tbody>
</table>

Table 4. Average values of human subjective evaluations for female speech samples. Index i of speech samples is the same as in Tab.2.

<table>
<thead>
<tr>
<th>i</th>
<th>( f_{i}^{\text{max}} )</th>
<th>Average students’ marks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original speech</td>
<td>Linear (5) transformation</td>
</tr>
<tr>
<td>2</td>
<td>8.36</td>
<td>1.80</td>
</tr>
<tr>
<td>3</td>
<td>6.86</td>
<td>1.48</td>
</tr>
<tr>
<td>4</td>
<td>6.17</td>
<td>1.61</td>
</tr>
<tr>
<td>5</td>
<td>—</td>
<td>1.72</td>
</tr>
<tr>
<td>6</td>
<td>2.2</td>
<td>1.71</td>
</tr>
<tr>
<td>7</td>
<td>3.99</td>
<td>1.57</td>
</tr>
<tr>
<td>8</td>
<td>8.66</td>
<td>1.59</td>
</tr>
</tbody>
</table>

Acknowledgement

This work was supported by MNiSW grant.

References


