MULTILINGUAL VOICE ANALYSIS: TOWARDS PROSODIC CORRELATES OF VOICE PREFERENCE

Horst-Udo Hain[1]*, Oliver Jokisch*[1] *and Luis Coelho*[2]

[1]*Laboratory of Acoustics and Speech Communication, TU Dresden, Germany*
[2]*Polytechnic Institute of Oporto, Portugal*
horst-udo.hain@mailbox.tu-dresden.de

Abstract: Finding an appropriate corporate voice is usally a time-consuming and laborious task. Additional constraints come into play if the recorded voice is the basis of a TTS system. Hence a corporate voice approach should retrofit conventional requirements like high intelligibility and naturalness of the synthetic speech signal.

The speaker selection is performed in several steps. After a preselection out of several candidates, the most promising speakers are recorded in a professional studio. These recordings together with resynthesised samples are then ranked by different listener groups.

The results of the subjective ranking are compared to objective measurements. First investigations show correlations between prosodic features and human judgement. Another conclusion is that although voice preference is a highly subjective decision and language-dependent, there are cross-language skills among native/non-native listeners and listeners without any skill in a specific language.

Kurzfassung: Ein wichtiges Kriterium bei der Beurteilung eines Sprachsynthesesystems ist die Qualität der Stimme, mit der es spricht. Deswegen wird die Auswahl des Stimmspenders mit großer Sorgfalt betrieben, was bisher auch immer mit entsprechend hohem Aufwand verbunden war, da sie auf der Grundlage der subjektiven Bewertung der Teilnehmer an den Hörtests erfolgt. Zum einen ist das Ergebnis aber stark von Art, Anzahl und Tagesform der Hörer abhängig, und zum anderen benötigt man für eine qualifizierte Einschätzung der Stimmen Muttersprachler. Der Prozeß zur Auswahl der Teilnehmer am Hörtest stellt also ebenfalls einen nicht zu vernachlässigenden Aufwand dar. Zusätzlich sollte noch bedacht werden, dass es durchaus eine Diskrepanz zwischen der Meinung der Sprachexperten und Akzeptanz der Anwender des Systems geben kann.

Um nun den Aufwand für diese Auswahlprozesse möglichst gering zu halten und weitestgehend zu automatisieren, sollen objektive Bewertungsmaße untersucht werden, bei deren Anwendung man Ergebnisse erhält, die mit denen der Hörtests vergleichbar sind. Diese Maße müssen automatisch aus dem Sprachsignal abgeleitet werden können und sprachenunabhängig sein. In der hier vorgestellten Arbeit werden in einer ersten Untersuchung F0 und Sprechrate für Datenbanken der Sprachen Deutsch, britisches Englisch, amerikanisches Englisch, Spanisch, Portugiesisch und Französisch analysiert und subjektiven Bewertungen gegenübergestellt. Im Ergebnis zeigt sich, dass die Werte der mittleren und maximalen Grundfrequenz sowie der Sprechrate mit der subjektiven Bevorzugung einzelner Sprecher korrelieren.

1 Introduction

An important task for the design of a TTS system is to find an appropriate voice that must be pleasant, natural and intelligible. The speaker selection process is usually time-consuming, and a lot of manual work is involved (cf. section 3.2). Additional problems occur if the developers are not familiar with foreign languages they have to deal with. Then they are reliant on the judgment of mother tongue experts who often do not have experience in speech processing.

Therefore the aim of the investigation described in this paper is to find objective criteria or cues that allow for an automated rating of different speakers, at least for a preselection to reduce the number of candidates to an amount that is manageable in a given time. Only those criteria come into consideration which can be automatically derived from the speaker database, such as fundamental frequency or speaking rate.

The results presented here are based on an analysis of recordings performed within a cooperation between Siemens AG, Munich, and TU Dresden for the creation of new voices for an embedded version of the multi-lingual TTS system "Papageno". In this project, amongst others, voices for German, UK and US English, French and Spanish have been recorded at TU Dresden laboratories [4].

The paper is organized as follows: section 2 explains the recording of the speaker databases, section 3 describes the prosodic correlates that have been analyzed, in section 4 the obtained results are discussed, followed by section 5 where some conclusions are drawn.

2 Speaker database and listening tests

2.1 Database for speaker selection

2.1.1 Requirements

Goal for the speaker selection was to find a voice for speaker recordings that could be used to create a concatenative (based on diphones) TTS system for embedded platforms. The supported languages were Dutch, UK and US English, French, German, Italian, and Spanish.

The speaker that will be finally selected must fulfill a number of requirements. In general, the voice has to be intelligible, natural and pleasant. A special demand is that it must be suitable for all the processing steps that are involved in speech synthesis. The voice quality (F0, jitter) must be sufficient and allow for good results even after compression or codecs (e. g. adaptive multi-rate, AMR) are applied. The speaker needs to have phonetic and also prosodic abilities (preferable a professional or semi-professional speaker) and should have experience in speaking a long time (about 4 hours per session) without any degradation of the voice quality (e. g. a teacher, actor, newsreader). An additional requirement is the use of a female voice. For all languages at least five speakers have to be recorded for the test database.

2.1.2 Multilingual corpus

The recordings contain up to 6 phonetically balanced sentences or phrases from a fairy tale to test the prosodic abilities, and about 50 single "carrier words" for a diphone test synthesis. Based on these carrier words, a small diphone corpus is created. The recorded sentences are then re-synthesised by the help of this corpus. Prosodic contours are mapped from the original sentences to assure that the synthetic sentences are not degraded by artificial prosodic artefacts. By these means, the expected voice quality of the real-world TTS system that is based on the recorded voice also has an influence on the speaker selection process.

2.2 Recording procedure and technical settings

The recordings took place in a studio with less reverberation (RT60 = 200 ms) using the large diaphragm microphone BPM Studiotechnik CR73-II including pop protector. The sampling rate was set to 44.1 kHz (mono channel) and a dynamics of 16 bit was provided. Additionally, the synchronized laryngograph signal (Lx Processor) was recorded for a potential precise analysis of the fundamental frequency.

2.3 Listening tests procedure

The voice preference was evaluated by pair comparison tests with 2 - 3 different test stimuli of 5 - 8 speakers (random selection), in total 40 - 72 phrases.
Two listener groups were involved: (1) about 20 native speakers who were mainly not familiar with speech and language processing, and (2) about 10 non-native speech experts. The presented voice prompts were either generated from original recordings or from the output of a simulated synthesis (using the small diphone test corpus). Both, the original (broad band) signal quality and a degraded quality (by AMR codec which is typical for mobile communication applications). The listeners had to select the better sample, which got one credit point for each selection. These credits were summed up and used as a score for the listening preference.

3 Prosodic correlates of voice preference

The intention of this survey is to find acoustic parameters that are correlated with the ranking obtained by subjective listening tests. Therefore at first such listening tests have to be performed and analyzed regarding their relevance and reliability. Then objective parameters are evaluated and compared to the results of the subjective tests.

3.1 Subjective tests

Concerning subjective tests, two different criteria are of interest. On one hand there is the question how different the rankings of native and non-native listeners are, and on the other hand the influence of the expert level has to be considered.

3.1.1 Differences between native and non-native listeners

The speaker selection described in section 2 for the different languages was performed by a subjective ranking of different groups of listeners: on one hand native and non-native speakers, and on the other hand listeners who are or are not familiar with speech processing technology. The question is now, how reliable and comparable these rankings are.
Figure 1 depicts the results presented in [4] for the selection of the UK English speaker out of six candidates. In most cases, the non-native listeners also preferred the candidate which received the highest rank by the native listeners. In some cases, the opinions between younger and older natives differed more than between natives and non-natives. Similar results were achieved for the other languages.

3.1.2 Influence of expert level

The effect of experience levels on voice quality ratings was studied in [2]. Ratings from speech and language therapists specialized in voice with at least 2 years experience are compared with those of final year speech and language therapy students. In total 14 parameters like breathiness, roughness and monotony as well as pitch or loudness were investigated. An important

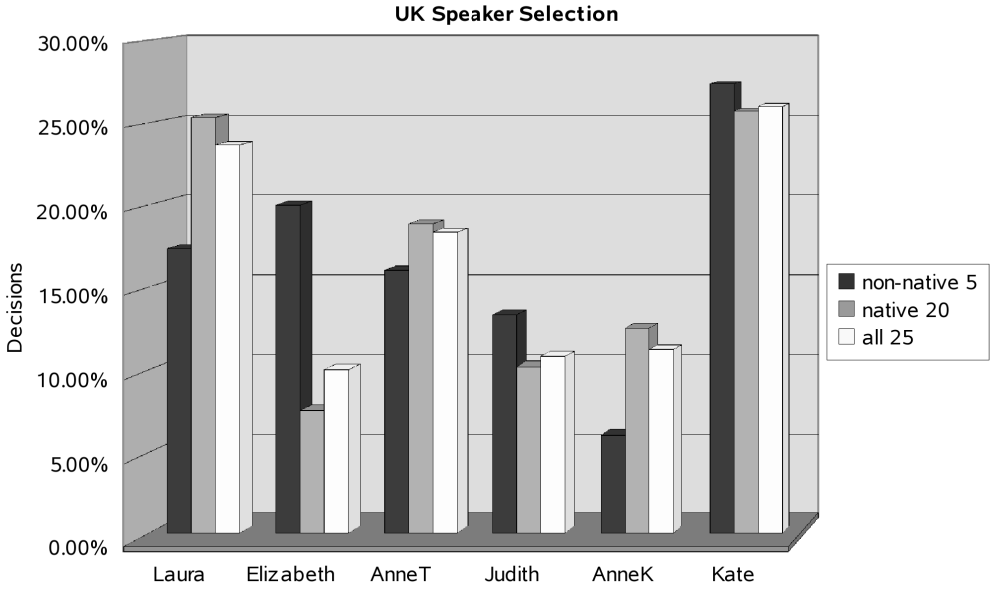

Figure 1 - Rankings of UK speaker selection.

basic condition is that only those perceptual labels should be used for a comparison with acoustic parameters which have a reliable judgment by both listener groups. The conclusion of this article is "that perceptual strategies between more and less experienced listeners are not different, but rather that these listeners adopt different baselines during perceptual tasks".

3.2 Objective parameters

Result of the speaker selection process is a ranking of the candidates according to their preference by the listeners. The task now is to find acoustic parameters representing this ranking. The previous investigations show a relation between acoustic parameters and voice quality ratings.

3.2.1 Glottalization

In [3], the influence of glottalization is addressed for UK and US English, Spanish, Italian, Dutch, and Chinese. Three types of glottalization are distinguished: vowel initial, phrase final, and additionally for Chinese glottalization in connection with tone 3. Utterances of native speakers are judged by both native and non-native listeners. One outcome of this article is that the most preferred speakers employ frequent glottalization.

3.2.2 Prosodic parameters

There is a less amount of technically-oriented references about the speaker selection process or correlating objective parameters. A speaker selection process for European Portuguese is described in [1]. The assessment explained there compares the results of a subjective evaluation with objective tests based on acoustic parameters. This article also contains a detailed description how laborious such a selection process can be. The first stage was a national call for voice talents which had to fulfill a few profile requirements. They had to be female, have European Portuguese as mother tongue, having studied in Portugal up to university level, speaking standard European Portuguese, and they should have some radio or theater vocal experience. Out of 485 candidates, 74 were invited to send samples of their voices with the maximum quality they could produce. A subjective test was then conducted with 13 questions, based on the MOS scale, with listeners who were familiarized with speech processing technology. The 12 best

scored candidates were invited to a professional studio for recording a small text. This procedure guarantees that all voices are evaluated under identical conditions. In contrast to the 5 points MOS rating of the second stage, now an exclusive multiple choice questionnaire was performed where only the best voice for each attribute could be selected. The listeners for this test were not familiar with speech processing technology. The final ranking was obtained through the sum of votes each voice received during the survey.

Additionally an objective analysis was carried out to confirm the evidence provided by the subjective tests. For that purpose the acoustic parameters F0 (mean, maximum, minimum, range and standard deviation), energy (mean and standard deviation), speaking rate (in words per minute excluding pauses) and pausing rate (total duration of voice sample without pauses) were used. A comparison of the best ranked voices with respect to the acoustic parameters yields the following summary:

- Listeners prefer voices with a mean F0 between 186 and 206 Hz and dislike voices with mean F0 ranges under and over these values.
- They also prefer a low minimum F0.
- A high speaking rate combined with long utterance breaks is favoured.
- Energy seems to be of little influence.

The outcome of this survey is now applied to the databases which have been recorded for the speaker selection described in section 1.

3.3 Analysis of speaker selection data

According to the results of [1], for the databases described in section 2.1 the acoustic parameters F0 (minimum, maximum, mean and standard deviation), speaking and pause time, average speaking rate, and energy (total value, average and standard deviation) have been calculated. Table 1 depicts the most interesting values.

Ranking	1	2	3	4	5
F0 Max	353.3	349.1	352.2	358.4	363
F0 Min	52.68	47.11	49.58	52.74	45.66
F0 Mean	201.2	200.1	208.9	213.6	215.1
F0 SD	57.18	60.31	58.31	58.07	61.77
Spk rate	2.7	2.79	2.74	2.69	2.38

Table 1 - Average values for the speakers with the ranking from 1 to 5 in every language: F0 in Hz (maximum, minimum, mean, standard deviation), and speaking rate (in words per second).

This table contains the average values for the speakers of equal rank for the languages German, UK and US English, Spanish and French. For example, the average of the maximum F0 values for the highest ranked speakers for every language is 353.3 Hz. The following figures depict the distribution of the most significant parameters.

First and second speaker rank mark an interesting finding: The average F0 mean over all five languages (cf. Figure 2) is below the upper limit of 206 Hz that is given in [1]. The authors come to the same conclusion that higher values of the F0 mean lead to lower listening preferences. Similar results can be obtained for the maximum F0 (cf. Figure 3). The preferred average speaking rate is increasing from rank five to two (Figure 4).

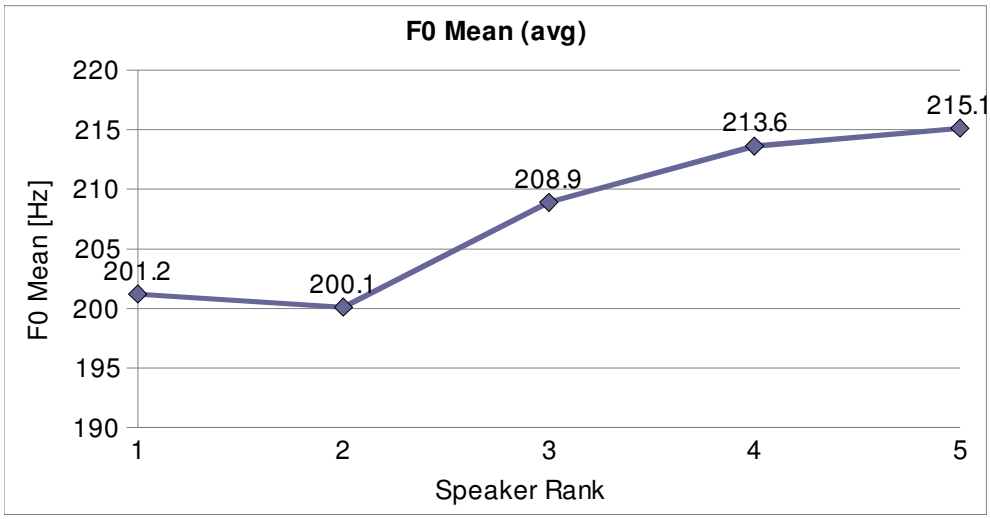

Figure 2 - F0 mean of the first five ranks.

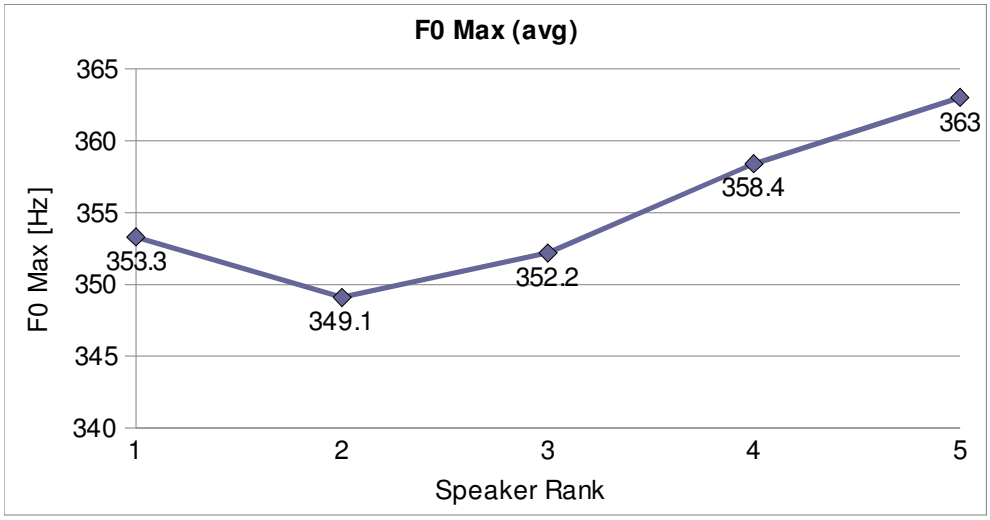

Figure 3 - F0 max of the first five ranks.

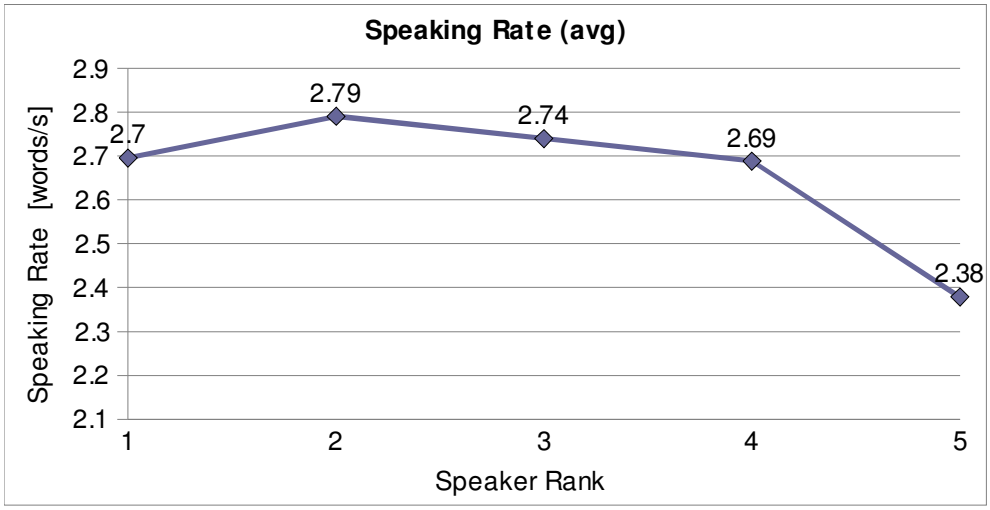

Figure 4 - Speaking rate of the first five ranks.

4 Discussion

Figures 2 to 4 do not show a monotonically increasing or decreasing behavior for the parameters from rank 1 to 5. The values of the first ranks are often rather similar. However, a tendency towards the lower ranks is evident, at least for these three depicted parameters. Other parameters like energy, minimum F0 or speaking time do not show a correlation between their values and the corresponding rank. Furthermore, the energy parameter (speech intensity) was normalized during the listening test.

Some of the results concerning the correlation between subjective tests and acoustic parameters presented in [1] could be confirmed, especially for mean F0 and speaking rate. Other values show a different behavior. One reason could be the small size of the databases of 5 - 6 speakers per language. Larger databases will be helpful for the further research.

5 Conclusions

This paper presented preliminary results which approve that there is a correlation between the voice quality ranking obtained by subjective listening tests and three acoustic parameters which can be automatically derived from the speaker databases. These parameters can therefore be used for an automatic preselection of promising speakers from a larger number of candidates. So far the languages German, UK and US English, French and Spanish have been analyzed. The obtained results are comparable to those of a different study about European Portuguese.

Further investigations will focus on the results obtained by different groups of listeners as young/old, native/non-native, and expert/non-expert regarding the speech processing technology. Additionally, further languages and potential correlates will be surveyed.

6 Acknowledgements

The authors would like to thank Daniela Braga from Microsoft Language Development Center for her valuable hints regarding the research topic.

References

[1] D. Braga, L. Coelho, F. G. V. Resende Junior, and M. S. Dias. Subjective and objective assessment of TTS voice font quality. In *International Conference Speech and Computer (SPECOM 2007)*, pages 306–311, Moscow, October 15–17 2007.

[2] C. de Bruijn and S. Whiteside. Effect of experience levels on voice quality ratings. In *Phonetics Teaching and Learning Conference*, London, August 2007.

[3] H. Ding, O. Jokisch, and R. Hoffmann. The effect of glottalization on voice preference. In *3rd International Conference on Speech Prosody*, volume II, pages 851–854, Dresden, May 2006.

[4] O. Jokisch, G. Strecha, and H. Ding. Multilingual speaker selection for creating a speech synthesis database. In *Workshop Advances in Speech Technology AST*, Maribor, Slovenia, 2004.