How to select a good voice for TTS

Sunhee Kim

Naver Labs, Naver Corporation, Korea

kim.sunhee@navercorp.com

Abstract

Even though the perceived quality of a speaker's natural voice does not necessarily guarantee the quality of synthesized speech, it is required to select a certain number of candidates based on their natural voice before moving to the evaluation stage of synthesized sentences. This paper describes a male speaker selection procedure for unit selection synthesis systems in English and Japanese based on perceptive evaluation and acoustic measurements of the speakers' natural voice. A perceptive evaluation is performed on eight professional voice talents of each language. A total of twenty native-speaker listeners are recruited in both languages and each listener is asked to rate on eight analytical factors by using a five-scale score and rank three best speakers. Acoustic measurement focuses on the voice quality by extracting two measures from Long Term Average Spectrum (LTAS), the so-called Speakers Formant (SPF), which corresponds to the peak intensity between 3 kHz and 4 kHz, and the Alpha ratio (AR), which is the lower level difference between 0 and 1 kHz and 1 and 4 kHz ranges. The perceptive evaluation results show a very strong correlation between the total score and the preference in both languages, 0.9183 in English and 0.8589 in Japanese. The correlations between the perceptive evaluation and acoustic measurements are moderate with respect to SPF and AR, 0.473 and -0.494 in English, and 0.288 and -0.263 in Japanese.

Index Terms: speech synthesis, speaker selection, voice quality, perceptive evaluation, acoustic measurement

1. Introduction

Given that a voice can generate a certain personality impression as reported in the area of psychology [1, 2], the choice of a good voice can give an impression of a better TTS system all else being equal except for the voice selected. This may be due to the nature of the voice, which necessarily cannot be all reflected in the system. It is reported that the quality of synthesized speech is not guaranteed by the perceived quality of the speaker's natural voice [3], however, it is required to select a number of candidates based on their natural voice before moving to the evaluation stage of synthesized sentences.

This paper describes a male speaker selection procedure for unit selection synthesis systems in English and Japanese based on perceptive evaluation and acoustic measurement of the speakers' natural voice. The rest of the paper is organized as follows. Section 2 presents related work performed in various fields, which determines the scope of the current paper. In Section 3, the overall speaker selection procedure for developing our TTS system is presented, which is then followed by the methods of perceptive evaluation and acoustic measurements. Section 4 includes the experimental results, which is followed by the discussion in Section 5. Section 6 then concludes the paper.

2. Related work

Before we address the problem of selecting a good voice for TTS, the notion of good voice needs to be defined. This topic has been investigated in different areas. From a perceptive point of view, the results are reported of interviews with radio employers and educators before they were trained to work in the areas of vocal training and clinical management of voice disorders [4]. According to the interviews, a good voice for radio performers is characterized as sounding easy-on-the-ear, including warmth, depth of pitch, clarity of speech, presence, animation, and liveliness. It is also noted that the content and personality are more significant than voice characteristics for radio performers. A summary on previous literature on communicative characteristics of radio performers is presented in [4]. The vocal attractiveness is also examined by making a distinction between voice cues (e.g., pitch, intensity etc.) and speech cues (e.g., non-fluencies, speech rate, etc.) as sources of interpersonal impressions [5]. According to [6], a good speaker is trustworthy, expressive, powerful and involved; whereas being insecure, hesitant, or monotonous lead to the opposite impression.

Research on the qualities contributing to the impression of a good speaker has been approached based on the investigation of the correlation between subjective ratings and acoustic measurements. It is reported in [6] that the F0 and duration data are correlated with the results of subject ratings while the speakers reached a high level of agreement on qualities contributing to their impressions. In [7], acoustic measurements which reveal substantial differences for speakers rated high and low are those related with F0 and duration features. According to [8], liveliness and speaking rate are correlated with general pleasantness of the voice based on their correlations with the features related with F0, Energy, and Spectral. To sum up, the impressions of a good speaker are found to be related to pitch, amplitude, speaking rate, and fluency.

The relationship between different emotions and acoustic features is also explored. It is reported that positive-activation emotions have a high mean F0 and energy as well as a faster speaking rate than negative-activation emotions [9]. It is also demonstrated how spectral tilt and pitch contour contribute to distinguishing emotions. On the other hand, charismatic voice is characterized by temporal and pitch structure of the voice, showing a strong correlation between the subject ratings of the charismatic statement and the acoustic features of pitch and intensity, along with the speaking rate and durational features [10]. Research on the automatic detection of speaking styles based on acoustic features has also been conducted by using a large database [8, 11, 12, 13].

Many studies, especially in the clinical voice research field, have approached the subject of overall voice quality through acoustic measurement. Previous quantitative research on the relation between perceived overall voice quality and several acoustic-phonetic measurements is reviewed in [14], reporting that four measures in sustained vowels (smoothed cepstral peak prominence, spectral flatness of residue signal, Pearson *r* at autocorrelation peak, and pitch amplitude) and three measures in continuous speech (signal-to-noise ratio from Qi, cepstral peak prominence, and smoothed cepstral peak prominence) are correlated with the overall voice quality. A similar approach has been applied to distinguish modal voice quality from abnormal voice quality, which is usually characterized by subjective measures such as Grade, Roughness, Breathiness, Aesthenia, Strain (GRBAS). In [14], it is also indicated that only three noise related parameters - voice turbulence index (VTI), noise harmonic ratio (NHR), and soft phonation index (SPI) - are significantly correlated with GRBAS perceptual voice analysis.

There are interesting studies on the voice quality of trained speakers along with that of trained singers, Speaker's Formant for trained speakers [15, 16] and Singer's Formant for trained singers [17]. Singer's Formant corresponds to the local maximum of energy or the cluster of formants between 2,300 and 2,900 Hz in male voices, while Speaker's Formant an increasing envelope peak between 3,150 and 3,700 Hz, confirmed statistically [16]. The Speaker's Formant is found in the region of the fourth formant, showing sonorous quality of the voice. It is reported that it increases gradually and is approximately 10 dB higher in professional male voices than in normal male voices at neutral loudness (60 dB at 0.3 min) [16].

The correlations between the acoustic measurements and the quality of synthesized speech is explored in terms of intelligibility, naturalness, and pleasantness for selecting a speaker for TTS, reporting that RMS energy for unvoiced speech, some long-term spectrum cepstral coefficients and pitch variation of female speakers are found to be correlated with subjective ratings [3].

This paper aims to examine the correlations between the results of perceptive ratings and the voice quality of the natural voice of candidates for the English and Japanese male speaker selection procedure in unit selection synthesis systems using two acoustic measurements proposed in [18]: Speaker's Formant (henceforth, SPF) and Alpha Ratio (henceforth, SPF) which corresponds to the lower level difference between 0 and 1 kHz and 1 and 4 kHz ranges. These measures focus on the voice quality, while F0, amplitude and speaking rate are related to liveliness or way of speaking [18]. The measures of noise or perturbation developed for use in the dysphonic population will not be considered in this paper, as the candidate speakers recruited for TTS are professional voice talents including broadcasters and voice-over artists, who do not show any abnormal voice qualities.

3. Method

3.1. Speaker selection procedure

NVOICE is the name of Text-to-Speech systems developed by NAVER Corporation in Korea, based on both unit selection synthesis and statistical parametric synthesis. This paper focuses on the unit selection based TTS system and the male speaker selection procedure of looking for a good voice in English and Japanese.

The entire process of speaker selection consists of three stages as presented in Figure 1. In the first stage, we are provided with voice recordings of 10-20 sample sentences of at least 20 professional voice talents (VTs) through an agency, among which 8 to 10 VTs are selected through perceptive and technical evaluation. If we do not find at least three VTs for the next stage of evaluation, we request more samples of other VTs. The perceptive evaluation of this stage is performed by the project members, including at least one native speaker of the language and a linguistic expert.

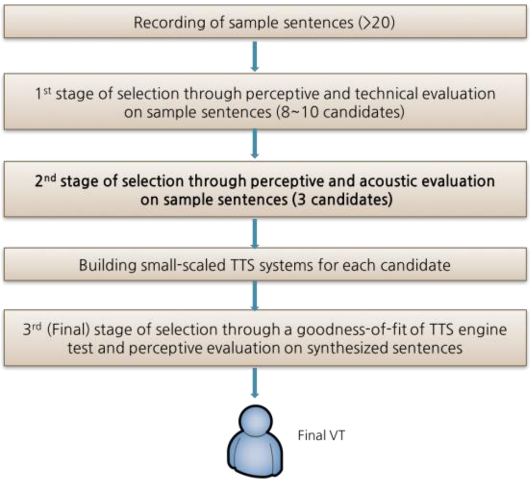
Recording of sample sentences (>20)
1st stage of selection through perceptive and technical evaluation on sample sentences (8~10 candidates)
2nd stage of selection through perceptive and acoustic evaluation on sample sentences (3 candidates)
Building small-scaled TTS systems for each candidate
3rd (Final) stage of selection through a goodness-of-fit of TTS engine test and perceptive evaluation on synthesized sentences
Final VT

Figure 1: *Speaker selection procedure*

The second stage consists of perceptive evaluation carried out by native speakers of each language and acoustic measurements of certain features. The final stage includes perceptive evaluation of synthesized sentences of each small-scaled TTS systems for each candidate, which is built on 500 to 1000 sentences of the three candidates selected in the second stage. A goodness-of-fit of TTS engine and perceptive evaluation by native speakers and linguistic experts is performed to choose the final speaker.

As our unit selection synthesis system needs a large speech database, consisting of about 30,000 sentences, the recording takes a few months. Thus, it is very important to select a speaker who has a healthy voice and can well manage his or her voice during the long recording period. This paper describes the second stage of the selection process, which consists of perceptive evaluation and acoustic measurement for selecting three candidates who supposedly have good voices for English and Japanese TTS. The correlations between perceptive evaluation and acoustic measurements are calculated and three candidates of each language are chosen for the final stage.

3.2. Perceptive evaluation

Sample recordings of eight English speakers and eight Japanese speakers are prepared for the evaluation. Eighteen sentences are recorded and presented to each rater, who is a native-speaker of the language. The sentences consist of eight phonetically rich sentences, four statements and four questions, and two sentences from news reports.

All raters are native speakers residing in their home country and recruited by the agency. All the speech files are converted into 16 kHz sampling rate from 48 kHz and normalized in amplitude. Each rater is asked to rate eight analytical factors of evaluation by using a 5-point scale. The analytical factors for evaluation are proposed as follows:

- Accent
- Clear pronunciation
- Correct pronunciation
- Liveliness
- Rhythm/Phrasing
- Coherence
- Pleasantness
- Trustworthiness

In addition to rating on analytical factors, each rater is also asked to rank the three best speakers, which is then employed to estimate the overall preference of each speaker.

3.3. Acoustic measurements

As is mentioned above, the acoustic measurement focuses on the voice quality by extracting two measures from Long Term Average Spectrum (LTAS), the Speakers Formant (SPF), the peak intensity between 3 kHz and 4 kHz, and the Alpha ratio, lower level difference between the 0-1 kHz and 1-4 kHz ranges. All processes are conducted by using Praat.

LTAS, denoting the Logarithmic spectral density as a function of frequency, is expressed in dB/Hz, which is usually computed over voiced speech segments. In order to extract the SPF from speech files, they are first resampled into 16 kHz from 48 kHz. Then, the LTAS values of each file are extracted with the maximum frequency range of 8 kHz and 100 Hz of bandwidth, resulting in 80 values. Each SPF, the maximum intensity between 3 and 4 kHz, which in fact corresponds to the region of the fourth formant, is obtained from these LTAS values. The Alpha Ratio (AR), the lower level difference between the 0-1 kHz and 1-4 kHz ranges, of each file is calculated by subtracting the total energy between 1 and 4 kHz from the total energy under 1kHz, representing greater energy in the 1 to 4 kHz range.

The same sample speech files used for perceptive evaluation are the object of acoustic measurements. Thus, the total number of sentences subject to the acoustic measurement amounts to eighty in each language.

4. Experimental results

4.1. English speakers

The overall agreement between the raters is 0.254. The total score and the mean value of each speaker, along with the preference results of perceptual evaluation are presented in Table 1. Preference is calculated by how many times each speaker is ranked as one of the top three speakers without considering the order. As twenty raters choose three people, the total score of the preference is 60. There is a very strong correlation between the total score and preference, which is 0.918.

Table 1: *The total score, mean value and preference for English VTs*

	M9	M15	M21	M23
Total	693	673	709	717
Mean	4.33	4.21	4.43	4.48
Preference	9	4	8	13
	M26	M29	M30	M40
Total	600	730	593	684
Mean	3.75	4.56	.71	4.28
Preference	2	14	1	9

Table 2 shows the correlation of each rater's score with the total score. The average correlation between each rater's score and the total score of all raters is 0.643. The correlation between the total score and the total score of all female raters is higher (0.690) than that of all male raters (0.596).

Table 2: C*orrelation of each rater's score with the total score of all raters (English).*

Female Rater	Correlation	Male Rater	Correlation
RF16	.960	RM03	.932
RF08	.897	RM17	.856
RF04	.849	RM01	.822
RF12	.819	RM10	.802
RF05	.735	RM11	.784
RF18	.616	RM06	.678
RF09	.615	RM13	.399
RF07	.562	RM02	.267
RF20	.440	RM14	.213
RF15	.405	RM19	.202

The correlation of each analytical factor of evaluation with their total score is shown in Table 3. The average correlation is 0.813.

Table 3: *Correlation of each analytical factor of perceptive evaluation (English)*

Trustworthiness	.975
Coherence	.952
Pleasantness	.920
Clear pronunciation	.910
Rhythm/Phrasing	.871
Liveliness	.749
Accent	.623
Correct pronunciation	.502

The average SPF and AR of each English speaker are presented in Table 4. The average SPF is 14.0 DB (SD: 5.9), and the AR -138.4 DB (SD: 103.1). The correlation between the total score and SPF is 0.473, while that of AR is -0.494.

Table 4: *The average SPF and AF of each speaker (English).*

	M9	M15	M21	M23
SPF	14.9	17.9	22.1	14.9
AR	-140.8	-197.7	-200.3	-287.4
	M26	M29	M30	M40
SPF	13.7	8.4	7.3	12.9
AR	-97.5	-17.9	23.3	-189.2

4.2. Japanese speakers

The overall agreement between the raters is 0.314. The total score and the mean of each speaker along with the preference results of perceptual evaluation are presented in Table 5. The average correlation between the total score and preference is 0.860.

Table 5: *The total score, mean value and preference score for Japanese VTs*

	M3	M6	M7	M10
Total	693	547	654	592
Mean	4.33	3.42	4.09	3.70
Preference	15	3	13	4
	M11	M15	M16	M20
Total	476	606	632	600
Mean	2.98	3.79	3.95	3.75
Preference	2	4	7	6

The correlation of each rater to the total score is shown in Table 6. The average correlation between each rater's score and the total score is 0.634. Unlike the American raters, the female raters show lower correlation (0.590 in average) than the male raters (0.679 in average).

Table 6: C*orrelation of each rater's score with the total score (Japanese).*

Female Rater	Correlation	Male Rater	Correlation
RF03	.979	RM05	.952
RF02	.871	RM06	.954
RF01	.827	RM04	.886
RF05	.738	RM10	.854
RF07	.714	RM01	.734
RF04	.614	RM07	.718
RF06	.611	RM08	.683
RF09	.521	RM03	.478
RF08	.308	RM09	.389
RF10	-.282	RM02	.153

The correlation of each analytical factor of evaluation with the total score is also calculated as shown in Table 7. The aver- age correlation is 0.933.

Table 7: *Correlation of each analytical factor of perceptive evaluation (Japanese)*

Liveliness	.986
Clear pronunciation	.982
Trustworthiness	.968
Accent	.953
Coherence	.940
Correct pronunciation	.934
Pleasantness	.897
Rhythm/Phrasing	.804

The average SPF and AR of each English speaker are presented in Table 8. The average SPF is 16.5 DB (SD: 4.6), and the AR -182.9 DB (SD: 97.2). The correlation between the total score and SPF is 0.288, while that of AR is -0.263.

Table 8: *The average SPF and AR of each speaker (Japanese).*

	M3	M6	M7	M10
SPF	18.1	24.5	18.2	19.6
AR	-171.8	-314.2	-186.2	-216.1
	M11	M15	M16	M20
SPF	8.1	18.2	14.4	19.2
AR	-1.7	-251.2	92.6	-229.2

5. Discussion

The overall agreement between the raters is 0.254 in English and 0.314 in Japanese respectively, which is not high due to the great amount of subjectivity, as reported in earlier studies [6], [10], [8].

The perceptive evaluation results show that there exit a very strong correlation between the total score and the preference in both languages, 0.918 in English and 0.860 in Japanese. We do not use the overall rating for each speaker, instead replacing it with the total score of analytical ratings. Furthermore, we have added another evaluation measure, preference, which is calculated by the number of times that each speaker is ranked as one of the top three speakers without considering the order. When the total score and preference are strongly correlated, the preference by itself can be used effectively.

The correlations between the total score and the sub-total score of each speaker obtained from each rater is 0.643 in English and 0.634 in Japanese. They are higher in female raters for English, while they are higher in male raters in Japanese. For each language, there is one rater who shows low correlation (under 0.25) and one female rater shows a negative correlation in Japanese. Despite the low overall agreement between the raters, this result may contribute to confirm the effectiveness of the procedure.

The correlation between the total score and each analytical factor of perceptive evaluation is also high in both languages, 0.813 in English and 0.933 in Japanese, despite the different ordering of the factors. It is interesting that while almost all the factors are highly correlated to the total score in Japanese, and that correct pronunciation is least correlated with the total score (0.502) in English. This result confirms that these factors can be used for further perceptive rating of natural voice, while intelligibility and naturalness are often used for the evaluation of synthesized speech.

The correlations between the total score of perceptive evaluation and the two acoustic measures are moderate both in English, 0.473 with SPF and -0.494 with AR, and in Japanese, 0.288 with SPF and -0.263 with AR. The results conform to those in [18]. In order to secure a large number of units available with varied prosodic and spectral characteristics for developing a TTS of natural-sounding quality, we employ a script for a large speech database, which amounts to at least 30,000 sentences. The results show that the SPF and AR are efficient in finding a speaker with a good voice, which can stay healthy until the completion of the long recording sessions.

The t-test results of SPF and AR between English and Japanese show that the two languages are not significantly different in terms of SPF and AR (p=0.159 for SPF and p=0.390 for AR).

6. Conclusions

This paper describes a male speaker selection procedure for unit selection synthesis systems in English and Japanese based on perceptive evaluation and acoustic measurements of the speakers' natural voice. A perceptive evaluation is performed on eight professional voice talents of each language. A total of twenty native-speaker listeners are recruited in both languages and each listener is asked to rate on eight analytical factors by using a five scale score and rank three best speakers. Acoustic measurement focuses on the voice quality by extracting two measures from LTAS, SPF and AR.

Despite the low overall agreement between the raters due to the great amount of subjectivity, which triggers moderate correlations between the total score and the sub-total score of each speaker, the results show a very strong correlation between the total score and the preference in both languages. The high correlation between the total score and each analytical factor of perceptive evaluation confirms that the proposed analytical factors can be used for further perceptive rating of natural voice. Even though the correlations between the perceptive evaluation and acoustic measurements are moderate with respect to SPF and AR, the results still show that these acoustic measurements contribute to estimate the voice quality of natural voice along with perceptive evaluation and other acoustic features such as pith, energy and speaking rate.

Finally, among the eight candidates in each language, the top three speakers, M21, M23, and M29 for English, M3, M7 and M16 for Japanese, based on their total scores and preference values, are selected for the final stage of perceptive evaluation of sentences which are synthesized from each small-scaled TTS system for each candidate.

7. Acknowledgements

We would like to thank the native-speaker raters who participated in the perceptive evaluation of English and Japanese.

8. References

[1] P. McAleer, A. Todorov, and P. Belin, "How do you say 'Hello'? Personality impressions from brief novel voices," *PloS one*, vol. 9, no. 3, e90779, 2014.

[2] G. W. Allport and H. Cantril, "Judging personality from voice," *The Journal of Social Psychology*, vol. 5, no. 1, pp. 37–55, 1934.

[3] A. K. Syrdal, A. Conkie, and Y. Stylianou, "Exploration of acoustic correlates in speaker selection for concatenative synthesis," in *ICSLP*, 1998.

[4] S. Warhurst, P. McCabe, and C. Madill, "What makes a good voice for radio: perceptions of radio employers and educators," *Journal of Voice*, vol. 27, no. 2, pp. 217–224, 2013.

[5] M. Zuckerman and R. E. Driver, "What sounds beautiful is good: The vocal attractiveness stereotype," *Journal of Nonverbal Behavior*, vol. 13, no. 2, pp. 67–82, 1989.

[6] E. Strangert, "What makes a good speaker? Subjective ratings and acoustic measurements," in *Proceedings from Fonetik 2007: Speech, music and hearing, quarterly progress and status report, TMH-QPSR*, vol. 50, pp. 29–32, 2007.

[7] E. Strangert and J. Gustafson, "What makes a good speaker? Subject ratings, acoustic measurements and perceptual evaluations," in *INTERSPEECH*, vol. 8, pp. 1688–1691, 2008.

[8] S. Mariooryad, A. Kannan, D. Hakkani-Tur, and E. Shriberg, "Automatic characterization of speaking styles in educational videos," in *Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4848–4852, 2014.

[9] J. Liscombe, J. Venditti, and J. B. Hirschberg, "Classifying subject ratings of emotional speech using acoustic features," In *INTERSPEECH*, 2003.

[10] J. B. Hirschberg and A. Rosenberg, "Acoustic/prosodic and lexical correlates of charismatic speech," in *INTERSPEECH*, pp. 513–516, 2005.

[11] F. Burkhardt, B. Schuller, B. Weiss, and F. Weninger, "Would you buy a car from me? - On the likability of telephone voices," in *INTERSPEECH*, pp. 1557–1560, 2011.

[12] S. Gonzalez and X. Anguera, "Perceptually inspired features for speaker likability classification," in *ICASSP*, pp. 8490– 8494, 2013.

[13] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. Van Son, F. Weninger, F. Eyben, T. Bocklet *et al.*, "The Interspeech 2012 Speaker Trait Challenge," in *INTER SPEECH*, 2012.

[14] Y. Maryn, N. Roy, M. De Bodt, P. Van Cauwenberge, and P. Corthals, "Acoustic measurement of overall voice quality: A meta-analysis," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2619–2634, 2009.

[15] I. V. Bele, "The speaker's formant," *Journal of Voice*, vol. 20, no. 4, pp. 555–578, 2006.

[16] T. Nawka, L. C. Anders, M. Cebulla, and D. Zurakowski, "The speaker's formant in male voices," *Journal of Voice*, vol. 11, no. 4, pp. 422–428, 1997.

[17] J. Sundberg, "Articulatory interpretation of the singing formant," *The Journal of the Acoustical Society of America*, vol. 55, no. 4, pp. 838–844, 1974.

[18] S. Warhurst, P. McCabe, E. Yiu, R. Heard, and C. Madill, "Acoustic characteristics of male commercial and public radio broadcast voices," *Journal of Voice*, vol. 27, no. 5, pp. 655–e1, 2013.