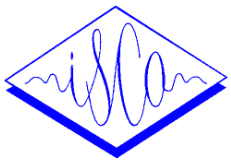

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

# VOICE QUALITY OF SYNTHETIC SPEECH: REPRESENTATION AND EVALUATION

***Louis C.W. Pols***

Institute of Phonetic Sciences / IFOTT, Univ. of Amsterdam, The Netherlands
tel: +31 20 525 2183 fax +31 20 525 2197 e-mail: pols@fon.let.uva.nl

## ABSTRACT

In most present-day rule synthesis systems (whether allophone-based or using concatenative units), the voice quality is generally limited to one voice and one speaking style. Although by now more knowledge is gathered about how to produce natural-sounding female voices, how to include some emotional elements in synthetic speech, and how to produce more acceptable prosody, a controlled and optimized voice quality of synthetic speech is so far much more a research goal than a reality.
This does not preclude of course good use of present-day synthetic speech for specific applications. But even under those conditions, the presently available methods to evaluate speech quality, barely touch on voice quality. Most emphasis is on phonemic quality, which is expressed in phoneme, word, or sentence intelligibility measures, or on very global measures such as overall quality, naturalness, or acceptability scores, generally collected via scale judgments.
We present some possibilities for diagnostic and functional evaluation of the voice quality of synthetic speech for specific applications.

## I. INTRODUCTION

*Voice quality*, at least in my interpretation for the present introductory paper, has to do with the segmental and supra-segmental characteristics of the voice source in (manipulated) natural, coded, or rule-synthesized speech.
• The *presence* of the voice source decides whether a speech segment will be voiced or unvoiced.
• The *timing* of the voice source defines the pitch level as such (male, female, child) as well as the pitch variations (declination; accent lending; phrase-final marking; expression vs. question; micro intonation), whereas it also influences speaking rate and rhythm.
• The *strength* of the voice source defines local (stressed syllables) and overall energy and loudness.
• Finally the *fine structure* of the single pitch periods of the voice source, or more specifically of the glottal flow, decides upon the way a voice sounds. This concerns *voice quality* in its narrowest sense.
The interplay of these characteristics also decides upon speaking style and emotion, although not exclusively since f.i. also pronunciation, dialect, and word choice determine whether a voice is will sound friendly or authoritative.

In my search for voice quality evaluation methods I will not limit myself to rule-synthesized speech. I would also like to involve the use of manipulated natural speech (e.g. key-words in carrier phrases; the use of techniques like PSOLA for local pitch and duration manipulations; voice conversion), analyzed-resynthesized or coded speech (allows for efficient storage, speech store-and-forward, and ciphering), or rule-synthesized speech from an (annotated) text or from a to-be-processed concept. In all instances the speech signal as a whole, or the voice source more specifically, may sound unnatural.
Also for natural voices it is sometimes desirable to evaluate its quality, f.i. for a singer in training or to control the treatment of a pathological voice. But there certainly is more reason for evaluating synthetic voices. In my opinion there are various types of questions one then can ask:

- does this artificial voice sound like an acceptable male, female, or child voice?
- does this voice sound like the intended voice of Mr. X or Mrs. Y? This includes the situation where one tries to adapt a synthetic voice to a given natural voice.
- does this voice have the intended characteristics in terms of speaking rate, rhythm, emotion, style, and perhaps otherwise?
- how would you specify the characteristics of this voice, f.i. on a number of semantic scales?

In terms of the methodology of voice quality evaluation, there is frequently one complicating factor, which is the fact that any test word or test sentence also has pronunciation characteristics and a meaning, from which one should like to abstract. This can be done, on the one hand, by using nonsense utterances or re-iterant speech, or on the other hand, by training the subjects to concentrate on voice characteristics alone. The type of tests one can think of are:

- speaker identification or verification (whose voice is this)
- (scale) judgment of voice characteristics (e.g. acceptability, naturalness, noisiness; which emotion is realized; use of the vocal profile analysis (VPA) protocol)
- comparison for similarity or preference judgment
- matching or adjustment (e.g. to equal prominence)
- diagnostic questions
- communicative, functional tests

As far as I know, very few systematic tests so far have been performed on the voice quality of synthetic speech. Probably closest comes the research towards a better sounding female voice [17, 18, 21], whereas also emotion starts to get some attention in speech synthesis [1, 5, 27, 32]. Intonologists have tried to optimize the features for sentence accent [10, 11, 30, 31] and so have people that are working on adequate spoken output in dialogue systems [3, 15]. Research on voice conversion [9] may give us some ideas, whereas also research on voice quality of pathological voices could be of interest [24, 25, 26]. Finally we may find inspiration in research on speaking styles ([12, 14], Barcelona ETRW), speaker characteristics (Edinburgh ETRW), prosody (Lund ETRW), synthesis in general (ETRW in Autrans and New Paltz, NY), and in specific research projects such as the BRA projects Speech Maps and Accor or the BRA Working Group VOX (analysis and synthesis of speaker characteristics).

## II. SEVERAL EXAMPLES

The probably most advanced rule-based synthesizer with good capabilities for voice source manipulations is the GLOVE synthesizer with a search-rule system [6, 7]. This system was for instance used to study the optimal settings of the LF-model glottal parameters [13], for seven female voices that were all judged normal by a speech therapist but more or less thin, breathy, dark and sonorant, etc. and three male voices. One cannot simply extract the relevant parameters from natural speech and one cannot simply derive unique relations for various voice qualities and speaking styles, so the optimization is frequently a process of trial-and-error. In the laboratory-development phase this is acceptable, however, once it comes to an operational system or to a formal product evaluation, one would like to have means also to specify a specific voice quality as well as the system's control capabilities. Below we give several examples of evaluation methods that might become such tools.

### 2.1 Speaker identification

Especially in low-bitrate speech coding it is not at all straightforward that the speaker characteristics are maintained after coding. LPC-resynthesis with a pulse source will almost certainly loose its speaker voice characteristics, although most timing aspects will remain. We showed one way to study speaker identifiability [4] by presenting first only very short fragments of speech and later on longer and longer passages until the speaker is correctly identified from a large potential group of speakers. The German Verbmobil project [33] is one of the first examples in which synthetic speech output is supposed to be tuned according to the intended speaker characteristics.

### 2.2 Identifying specific attributes

Specific attributes are identified when a listener for instance has to indicate the emotion expressed in a specific synthetic sentence (*neutral*, *joy*, *boredom*, *anger*, *sadness*, *fear*, or *indignation*). Vroomen et al. [32] studied that by manipulating only pitch and duration in a rule-based way.

### 2.3 Scale judgments of voice characteristics

Childers and Lee [8] consider the three scales *naturalness* (human sounding), *breathiness* (related to incomplete glottal closure) and *hypo/hyperfunction* (strained or tense vocal quality vs. too little tension in the vocal folds resulting in a thin or lax voice quality) sufficient to judge the voice quality of their source model when one parameter at a time is systematically manipulated for four voice types: modal, vocal fry, falsetto, and breathy.

De Leeuw [26] starts with 22 bipolar semantic scales to describe normal and pathological natural voices reading aloud texts. Some 16 scales remain after data reduction, they describe a 6-dimensional perceptual space representing *abnormality*, *tempo*, *pitch*, *strength*, *articulation quality*, and *melodiousness*.

De Krom [24] indicates that more and more consensus is reached about a few 'core' voice quality dimensions that can be distinguished by the majority of listeners. For the evaluation of dysphonia the so-called GRBAS scales are proposed: *grade G* (deviance of the voice), *degree of roughness R*, *breathiness B*, *asthenicity A* (lack of power, voice weakness), and *strain S*. De Krom added in his experiments the scale *instability*.

The well-known Vocal Profil Analysis Protocol [25] uses four feature sets (Vocal Quality, Prosodic, Temporal Organisation, and Comments) with various subcategories to judge pathological voices. Scoring requires preliminary training with audio-taped material and a manual.

Kreiman et al. [23] performed detailed experiments about *roughness* ratings of male voices with voice disorders, and express a warning concerning inter- and intra-rater reliability. They propose the use of fixed external standards or 'reference voices'.

### 2.4 Similarity or preference judgments

Kempster et al. [20] use similarity judgments between pairs of voices to create a low-dimensional perceptual representation of dysphonic female voices. The three main dimensions appear to be *intensity*, *fundamental frequency*, and *perturbation*.

More relevant for optimal synthetic voice quality are probably the paired-comparison preference judgments as for instance used by Carlson et al. [7] to optimize the voice source specifications of a female voice.

### 2.5 Matching or adjustment

A recent experiment by Repp et al. [30] is a good example of the *adjustment or matching task*. Subjects were requested to adjust the second $F_0$ peak in a synthesized sentence (with two accent peaks realized by the so-called pointed hat pattern) such that it would match the first peak in prominence. The heights of both peaks as well as the overall declination rates were systematically varied. Actually the subjects performed also a *judgment task* in which they just had to indicate which of the two accented syllables was the more prominent one. Both experimental procedures showed that the second peak could be lower for having equal prominence, actually more so for higher first peaks, however, the exact relation with the declination rate was less clear.

### 2.6 Diagnostic tests

Diagnostic questions can be studied with any of the preceding methods, however, the paired-comparison preference judgment and the adjustment to equal whatever (e.g. prominence, noisiness, spectral balance, female voice quality) seem to be most practical.

### 2.7 Functional tests

The functional tests that I have in mind here not so much concern evaluating specific source or filter functions, but rather the communicative functionality of the whole system or specific sub-modules of it. This relates to questions such as:

- does the intended word sound as actually being in focus?
- is the speaking rate and pitch range of this text-to-speech system appropriate for extended use of proofreading Japanese newspaper manuscripts [19]?
- is this sentence accent realization adequate [2]?
- sounds this specific synthetic sentence realization as the appropriate answer to the preceding question?

This last problem was studied by Grice et al. [15] in a joint SAM/SUNDIAL pilot experiment in which two prosody algorithms were compared in simulated dialogues. Also the actual presentation of the leading sentence (only on screen, or naturally spoken by a male or a female speaker) was varied, but differences between these three protocol types did not reach significance. However, for most sentences the annotated prosody algorithm was preferred over the default algorithm.

At a more basic level, Bladon [3] proposed a *Prosodic Assessment Suite* with 400 test items (all requiring Pass/Fail judgments) divided in 19 groups, each group testing some aspect of English prosody. Unfortunately, the approach is not general enough to be uniformly applicable. For instance, the tests in his group 7 relate to the properties of rising intonation typical of 'question-word' interrogatives, whereas questions can also be realized otherwise. Part of his tests will only produce positive results once (future) synthesizers actually embody sophisticated syntactic parsing and some semantic knowledge.

This diagnostic (or glass box) approach is the opposite of the black box approach in which the overall performance, in an actual situation, as judged by naive users, is determined. From the user's point of view this might indeed be the only thing that matters.

The ITU-TSS (International Telecommunication Union - Telecommunication Standardization Sector) [16] takes this extreme viewpoint by promoting a black box approach in evaluating the subjective performance assessment of speech voice output devices over the telephone. Users are asked to give their impression on multiple opinion scales. In addition to two overall quality scales (*overall impression* and *acceptance*) six other scales are proposed (*listening effort, comprehension problems, articulation, pronunciation, speaking rate,* and *voice pleasantness*). The judgments are based on 10 to 30 seconds of speech per message [22]. The scores leave little room for diagnostic evaluation, and do not allow one to tell, for instance, whether the text pre-processing module, or the voice source, or the unit concatenation led to a specific low pronunciation score.

## III. DISCUSSION

From the examples given above it will be clear that quite a few methods are available and actually have been used in the various laboratories to evaluate, compare, and improve the voice quality of speech synthesizers. The clinical practice of evaluating normal and pathological voices has added to this inventory of methods.

However, this has not yet resulted in a high-quality, rule-based, voice quality control system fully integrated in a rule synthesizer. So there is still a need for good ideas, well-structured (prosodic) databases to extract regularities, and good evaluation methods, to improve further the voice quality of synthetic speech.

For future use I can see a need for at least three different types of voice quality evaluation methods.

- First of all the *linguistic processing* should be evaluated in order to see if the modules make the proper text interpretations, not just in terms of grapheme-to-phoneme conversion, word stress assignment, sentence accent, and duration rules, but also in terms of required speaker and voice characteristics, dialect, style and speaking rate. It might be almost impossible to extract all this from text input only, although a narrator can get quite far. However, in rule synthesis from concept, or in machine translation, much of this knowledge might be available in a properly coded form.
- A second level of evaluation concerns the actual *acoustical realization* of the required characteristics. Most laboratory tests are done at this level. This concerns, on the one hand, the modelling of the exact glottal flow form, the detailed realization of the $F_0$ contour, and all the spectro-temporal features. On the other hand there is the effect this has on the judged voice, speaker, style, and emotion characteristics of the realized synthetic speech.
- Finally there is the level of the *communicative, functional realization*. In sect. 2.7 a few examples of possible tests were presented. Once the synthesizers work properly at the symbolic level and at the acoustic realization level, or offer no more opportunities for aditional tuning, one could argue that a subjective category rating test of the ITU-TSS type is all we need, also when we are specifically interested in the voice quality.

However, in the development phase a suite of tests, studying specific details, will still be required [28, 29].

## IV. REFERENCES


[1] Abadjieva, E., Murray, I.R. & Arnott, J.L. (1993), "Applying analysis of human emotional speech to enhance synthetic speech", *Proc. Eurospeech'93*, Berlin, Vol. 2, 909-912.

[2] Bezooijen, R. van & Pols, L.C.W. (1989), "Evaluation of a sentence accentuation algorithm for a Dutch text-to-speech system", *Proc. Eurospeech'89,* Berlin, Vol. 1, 218-221.

[3] Bladon, A. (1990), "Evaluating the prosody of text-to-speech synthesizers", *Proc. Speech Tech'90*, 215-220.

[4] Boxelaar, G.W. (1986), "PB-word intelligibility and speaker identifiability of 5 medium band coders: A pilot study", *IFA Proc.* 10, 97-111.

[5] Carlson, R. (1991), "Synthesis: Modelling variability and constraints", *Proc. Eurospeech'91*, Genova, Vol. 3, 1043-1048.

[6] Carlson, R., Fant, G., Gobl, C., Granström, B., Karlsson, I. & Lin, Q. (1989), "Voice source rules for text-to-speech synthesis", *Proc. IEEE-ICASSP89*, Vol. 1, 223-226.

[7] Carlson, R., Granström, B. & Karlsson, I. (1990), "Experiments with voice modelling in speech synthesis", *Proc. ETRW on Speaker characterization in speech technology*, Edinburgh, 27-39.

[8] Childers, D.G. & Lee, C.K. (1991), "Vocal quality factors: Analysis, synthesis, and perception", *J. Acoust. Soc. Am.* 90(5), 2394-2410.

[9] Childers, D.G. & Wu, K. (1990), "Quality of speech produced by analysis-synthesis", *Speech Comm.* 9, 97-117.

[10] Collier, R. (1993), "On the communicative function of prosody: some experiments", *IPO Annual Progress Report* 28, 67-75.

[11] Collier, R., Zitter, A. de & Terken, J. (1989), "On the perceptual salience of melodical variations and its consequences for intonational synthesis", *Proc. Eurospeech'89*, Paris, Vol. 2, 108-111.

[12] Eskénazi, M. (1993), "Trends in speaking style research", *Proc. Eurospeech'93*, Berlin, Vol. 1, 501-509.

[13] Fant, G., Liljencrants, J. & Lin, Q. (1985), "A four-parameter model of glottal flow", *STL-QPSR* 4/1985, 1-14.

[14] Granström, B. (1992), "The use of speech synthesis in exploring different speaking styles", *Speech Comm.* 11(4-5), 347-356.

[15] Grice, M., Vagges, K. & Hirst, D. (1992), "Prosodic form tests" and "Prosodic function tests", *SAM final report.*

[16] ITU-TSS (1993), "Subjective performance assessment of the quality of speech voice output devices", *Draft Recommendation P.8S 3/1993, COM 12-6-E.*

[17] Karlsson, I. (1990), "Voice source dynamics for female speakers", *Proc. ICSLP'90*, Kobe, Vol. 1, 69-72.

[18] Karlsson, I. (1992), Modelling speaking styles in female speech synthesis", *Speech Comm.* 11, 491-495.

[19] Kasuya, H. (1993), "Significance of suitability assessment in speech synthesis applications", *IEICE Trans. Fundamentals*, Vol. E76-A, No. 11, 1893-1897.

[20] Kempster, G.B., Kistler, D.J. & Hillenbrand, J. (1991), "Multidimensional scaling analysis of dysphonia in two speaker groups", *J. Speech and Hearing Res.* 34, 534-543.

[21] Klatt, D.H. & Klatt, L.C. (1990), "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *J. Acoust. Soc. Am.* 87(2), 820-857.

[22] Klaus, H., Klix, H., Sotscheck, J. & Fellbaum, K. (1993), "An evaluation system for ascertaining the quality of synthetic speech based on subjective category rating tests", *Proc. Eurospeech'93*, Berlin, Vol. 3, 1679-1682.

[23] Kreiman, J., Gerratt, B.R., Kempster, G.B., Erman, A. & Berke, G.S. (1993), "Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research", *J. Speech and Hearing Res.* 36, 21-40.

[24] Krom, G. de (1994), *Acoustic correlates of breathiness and roughness. Experiments on voice quality*, Ph.D. thesis, University of Utrecht, 145 pag.

[25] Laver, J. (1991), *The gift of speech. Papers in the analysis of speech and voice*, Edinburgh Univ. Press, 400 pag.

[26] Leeuw, I. de (1991), "Perceptual evaluation of voice quality before and after radiotherapy of patients with early glottic cancer and of normal speakers", *IFA Proc.* 15, 109-120.

[27] Murray, I.R. & Arnott, J.L. (1993), "Towards the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion", *J. Acoust. Soc. Am.* 93(2), 1097-1108.

[28] Pols, L.C.W. & Jekosch, U. (1994), "A structured way of looking at the performance of text-to-speech systems", *Proc. ESCA/IEEE Workshop on Speech Synthesis*, New Paltz, NY.

[29] Pols, L.C.W. & SAM-partners (1992), "Multilingual synthesis evaluation methods", *Proc. ICSLP'92*, Banff, Vol. 1, 181-184.

[30] Repp, B.H., Rump, H.H. & Terken, J.M.B. (1993), "Relative perceptual prominence of fundamental frequency peaks in the presence of declination", *IPO Annual Progress Report* 28, 59-62.

[31] Strik, H. & Boves, L. (1992), "On the relation between voice source parameters and prosodic features in connected speech", *Speech Comm.* 11, 167-174.

[32] Vroomen, J., Collier, R. & Mozziconacci, S. (1993), "Duration and intonation in emotional speech", *Proc. Eurospeech'93*, Berlin, Vol. 1, 577-580.

[33] Wahlster, W. (1993), "Verbmobil: Translation of face-to-face dialogs", *Proc. Eurospeech'93*, Berlin, Vol. Opening and Plenary Sesions, 29-38.