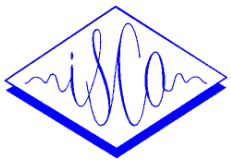

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

CONTROLLING VOICE QUALITY OF SYNTHETIC SPEECH

Inger Karlsson

Dept of Speech Communication and Music Acoustics, KTH,
Box 70014, S-100 44 Stockholm, Sweden

ABSTRACT

The paper will give an short exposé of some of the experiments performed to produce good and different voice qualities through the history of speech synthesis. The experiments have mainly been concerned with creating one or more natural-sounding voices for a specific speech synthesiser. Some examples of productions of emotive speech and of changes of voice qualities between for example normal and breathy voices are also mentioned. Emphasis will be on recent synthesis experiments performed at KTH.

INTRODUCTION

Through the years, people have tried to give their synthesisers more natural-sounding voices and also tried to produce different, mainly male, female and child, voices and different voice qualities. So far, these attempts have not been completely successful. Larger differences in voice quality, like sex differences, were produced already with the earliest synthesisers, even though the voice qualities were not very natural. With the developing of more sophisticated synthesisers, finer voice quality differences could be synthesised and the voices could be made to sound more natural. In combination with new sophisticated software analysis programs, this has opened up new possibilities for a larger variation of voices and voice qualities from speech synthesisers. The aim of this paper is to give an overview on what has been and is being done to produce good voice qualities with formant speech synthesisers.

Voice quality is in this paper treated as that which changes the acoustical impression from one speaker to another and for the same speaker, between different modes. It is also regarded as the difference between machine-like and natural-sounding speech synthesis. Accordingly, some factors pertaining to the durations of segments are included as well as vocal tract and subglottal pole and zero configurations and voice source variations.

Different voice qualities are used in speech to convey, among other things, different suprasegmental aspects, e.g., emphasis, phrase boundaries and also different speaking styles such as an authoritative or a submissive voice. Voice quality variations are also an important means of conveying extralinguistic information of various kinds, for example emotions, in ordinary speech. In this paper very little will be mentioned about when the different voice qualities are used, it will concentrate on experiments to produce and vary voice quality in synthetic speech.

DIFFERENT SYNTHESIS METHODS

In different synthesis methods different strategies can be used for varying voice quality. In diphone synthesis, perhaps the most wide-spread synthesis method today, some manipulation of spectral shape, F0 and duration can be made, which can give some small change in voice quality, see for example [19]. The normal way to change voice quality in diphone synthesis, though, is to record new speech material. Diphone synthesis will not be discussed further in this paper.

In formant synthesis, two different synthesiser configurations have been used, namely parallel and cascade synthesisers. The most important difference between the two for the production of different voice quality variations is, that in a parallel synthesiser, the voice source variations can be partly realised by varying the formant amplitudes. In a cascade synthesiser, the voice source is separated from the vocal tract. Most of the experiments discussed in this paper where performed on cascade synthesisers as these are, due to the separation of voice source and vocal tract, easier to manipulate and understand. For further discussions, see [16].

OVERVIEW

In the first synthesisers, it was only possible to change the first three formants, the amplitude and the fundamental frequency. Even so, experiments to create different voices were performed. Both G Fant using OVE I [6], and W Lawrence using PAT [23] tried to produce both male and female voices and W Lawrence also demonstrated an example of a "drunken speaker" [23]. This later example was produced by slowing the formant movements and reducing the vowels. The overall quality was not very good though, neither male nor female voice sounded natural.

With a later version of OVE, called OVE II, J Holmes produced a good copy of a male utterance while the copy of a female utterance was not as convincing [9]. In this synthesised utterances, the movements of the first three formants, durations, F0 movements and amplitudes were copied, but a simple impulse voice source was used to excite the synthesiser. Apparently, something more was needed to make it possible to produce high quality voices. Further experiments with male-female synthesis are discussed in [12].

Voice source experiments

An crucial factor deciding voice quality is voice source variations [18]. In the first synthesis experiments with different voice sources, copies of a natural utterance made by a male speaker was resynthesised using formants, durations and F0 variations measured from the speech signal and different voice sources [10,21]. The copy synthesis was then compared with the natural speech in perception tests to decide which type of voice source gave the best perceptual correspondence. For a parallel synthesiser, a copy of one natural voice pulse was preferred [10]. The same pulse was repeated through the utterance, only F0 was varied. In this case, some of the voice source variations could be modelled by amplitude differences for the different formants and by adding noise in the formant circuits. Rosenberg made similar experiments with a cascade synthesiser [21]. His results indicated that a polynomial voice source pulse containing one discontinuity gave the most natural voice quality. Both these experiments were using the same voice source pulse through the whole utterance. This is not sufficient to give a completely natural voice, as in real speech, the voice source varies considerably. The next step in the development of text-to-speech synthesisers was to include voice sources that could be controlled and varied.

One early, interesting example of a voice source model that could be changed by rule was the source constructed by Rothenberg et al. [22]. This source was used in conjunction with the OVE III synthesiser. The voice source was controlled by three articulatory parameters. The Loudness parameter was related to the subglottal pressure and mainly influenced the spectral tilt, the Tightness parameter influenced the skewness of the glottal pulse and the open quotient as well as the spectral tilt and Frequency decided the F0. Noise was generated differently depending on the peak glottal flow. This voice source gave a natural-sounding synthesis of different voices and voice qualities. The use of this voice source had to be discontinued as it proved to be very complicated to maintain and copy the original analogue voice source model.

The search for a good source model is continuing. The technique to copy a natural utterance have been used in many experiments. Only very few will be mentioned here, a comparison between some proposed voice source models can be found in [8].

Pinto et al. used copies of speech samples from male, female and child speakers and tested different voices sources, including source tract interaction in one of the sources, for the KLATT synthesiser KLSYN [20]. They found that a source with vocal tract interaction gave better results than a Rosenberg or the built-in impulse source. They also demonstrated phonation type changes by varying different voice source parameters.

Imaizumi et al. tested their own 7 parameter voice source model using copy synthesis of vowels uttered by males and females at different loudness and F0 [11]. The model produced good male synthesis, less good female.

OTHER QUALITY DECIDERS

Voice quality differences are not manifested solely in the voice source. The timing of different gestures, for example when voicing starts and ends, how fast formant transitions and fundamental frequency changes are made, or in which segments noise occur will all influence the perceived quality. An important factor in varying voice quality is, beside changing the shape of the glottal pulse, the adding of noise in voiced segments. Especially in breathy voice quality, the timing of the noise addition is crucial.

Copy synthesis and voice source variations has been used to imitate breathy voices. Childers et al. analysed the voice source for different voice types: modal voice, vocal fry, falsetto and breathy voice [5]. The results were resynthesised. They found that for a breathy voice, the noise should be amplitude modulated so that it occurred during 50 to 75% of the voice pulse and that the noise source should be located near the point of maximal closure. Kasuya et al. copied breathy voices using a Klatt synthesiser and found that to achieve a breathy impression, the noise amplitude should be amplitude modulated with the maximum amplitude coinciding with glottal opening [15].

In leaky voices, as well as in aspirated voiced segments, the coupling to subglottal cavities will be evident as extra spectral peaks and zeros in the spectrum and perceptible in the resulting speech. The occurrence of these poles-zeros is often dependent on coarticulation with adjacent phonemes and the frequency of occurrence is higher for leaky voices. This feature has been used by Klatt in copy synthesis of female voices [17], more about this below.

SYNTHESIS OF EMOTIONS

In many applications for speech synthesis, for example as speech prosthesis, individual voices and the possibility to express emotions and attitudes are important. Experiments to supply these possibilities for speech synthesis have been performed using both a complete text-to-speech system [1,2], and copy synthesis [4]. Editors for speech synthesis systems have been constructed, called HAMLET [1], and the Affect Editor [2]. Both these editors allow the user to produce different emotions by producing control parameter values for different aspects of F0, speech rate, loudness, pausing, precision of articulation and in the Affect Editor, some voice source characteristics. The parameter values were taken from the literature. A perception test using the Affect Editor gave very good results for the sad voice quality, slightly less for glad and disgusted, less again for scared and surprised and least for anger. For HAMLET, the angry, sad and fear emotions were most easily perceived.

The experiments carried out at KTH [4] consisted of copy synthesis of four different emotions: happy, neutral, sad and angry. The copied speech samples were recorded by a male actor and had been judged in a perception test to convey the intended emotions. Durations, F0 contours, intensities and formants were copied and also vocal fry. The synthesised copies were either produced by using all parameters from one emotion or by imposing durations and F0 contour of one emotion on another. Also here the sad voice was easy to recognise, 95%, and anger was some-

what harder while happy was much harder, 42%, and was often heard as angry. The mixed stimuli with sad durations and F0 contour were most often perceived as sad, while for the other mixed stimuli the percepts were more varied.

MODERN SYNTHESISERS

Two recent variants of formant synthesisers have created the possibility for the production of better and more varied voice qualities. These are the KLSYN88 [17] and the GLOVE synthesisers [3].

KLSYN88

This synthesiser is a development of the older KLSYN synthesiser. In the newer synthesiser, three different voice sources have been included. These are the old impulse source, the LF-source and a voice source model suggested by D Klatt. The voice source control parameters for this model are the open quotient, the spectral tilt and the skewness of the pulse, diplophonic double pulsing and a slow quasirandom drift of F0. Other important parameters for voice quality are that the first formant frequency and bandwidth can vary between the closed and the open part of the voice pulse. A pole-zero pair that can model the tracheal influence is also included [17].

The capacity of the synthesiser with the Klatt voice source was tested in the creation of good copies of several women's productions of reiterant speech, that is imitating an utterance by repetition the same syllable. It was necessary to mix in noise above 2 kHz in the vowels for some speakers to get good copies. Those speakers were perceived as having breathy voices. In a listening test, the synthesised and natural versions of [ʔɑ] and [hɑ] reiterant sentences from these female speakers were virtually indistinguishable.

GLOVE - THE KTH SYNTHESISER

The GLOVE speech synthesis model [4], is an improved version of the OVE III cascade synthesiser. In GLOVE, the LF-source [7], diplophonia (DI), F1 bandwidth open quotient dependency and noise/voice source interaction have been added. The LF source pulse is defined by four parameters: the shape parameters RG, RK, FA and the excitation energy EE. RG decides the balance between the first and second harmonic, a high RG gives more energy in the second harmonic. RK is decided by the skewness of the glottal pulse. A high RK value means that the first to second harmonic is comparatively strong. FA controls the spectral tilt, a higher FA gives more high frequency energy.

The noise/voice source interaction is specified by the parameters NA and NM. NA specifies the amount of glottal flow modulated noise added to the glottal pulse. NM specifies the degree of glottal flow modulation of the noise source. The higher formants, which was realised as a higher-pole correction in male speech produced with the OVE III model, are for the female voice experiments implemented as three separate formants (F5-F7), with constant relative distance. F5 can be changed by rule.

Naturalness

As a test of the GLOVE synthesiser, a crude copy of a female utterance was produced [3]. The segment durations and F0 curve was copied as well as the formant values at turning points. This resulted in 1-2 formant specifications per phoneme. The voice source was decided using inverse filtering for the mid point of each phoneme and in transitions between phonemes The number of data points was accordingly low and equivalent to the number of data points in text-to-speech synthesis. F0-modulated noise was added in some voiced consonants and the end of the sentence was made diplophonic by using the DI parameter. The GLOVE utterance was compared with the same OVE III produced utterance using the same F0, duration and formant settings. The GLOVE utterance sounded considerably more natural, even if there is still some difference between this and the natural utterance.

VOICE QUALITIES

Investigations of voice source variations in female voices with varying voice qualities have been performed, [13]. The data from these investigations were formalised and used in synthesising different voice qualities. In the synthesis experiment, only the voice source parameters were varied in a two-word utterance: [j'ɑ: aj'ø:]. The differences were discernible over earphones. The different qualities are described according to the deviation from the sonorant voice.

Sonorant - non-sonorant:

A non-sonorant voice quality have lower FA values in vowels than a sonorant voice quality, especially in the more open vowels, while FA for consonants is more equal for the two voice qualities. The FA range for the non-sonorant voice is considerably lower than for the sonorant voice, and the highest FA value for the non-sonorant voice is only about 20% of the highest FA value for the sonorant voice.

Voice quality: Strained:

A more tight and strained voice quality is characterised by a considerably higher RG and a lower RK for all speech segments compared to a normal, sonorant voice quality. FA for the strained voice is often equal to FA for a sonorant voice quality, but often FA does not fall towards the end of an utterance as it does in a sonorant voice. The high FA at the end of a phrase is combined with a creaky voice source, produced by DI, for the strained voice quality. When the RG parameter is raised very much, an impression of shouting is achieved.

Voice quality: Breathy - non-breathy:

As was shown in earlier mentioned experiments [5,15], the noise in voiced segments need to be modulated by the voice source frequency to be perceived as part of the articulation. The intensity of noise excitation is not constant through all the segments but stronger for certain segment types, typically consonants and transitions, and often also for the end of a phrase. When an equal amount of noise was added to all segments, it was heard as a noisy recording, even that the noise was modulated by the glottal flow and shaped by the same formant filters as the voiced source.

SYNTHESIS OF A FEMALE VOICE BY RULE

At present, the GLOVE synthesiser is being provided with a female voice that is intended for inclusion in the

KTH text-to-speech system [14]. The emphasis so far has been on the segmental level. To get better consonants, extensive use have been made of the LF-source and the pole/zero pair in the vocal branch. The pole is implemented as a formant filter. The zero consists of an inverse formant filter. Both are controlled by specifying mid frequency and bandwidth. When the pole/zero pair is not needed, both are placed at the same frequency and given a very low Q-value and will thus cancel each This pole-zero pair has so far been for consonants and vowel-consonant transitions. In the future, we intend to test its usefulness for voice quality variations between speakers as well.

Data for the female voice segments have been collected using mainly inverse filtering technique. This has given us formants, formant bandwidths, voice source parameters as well as pole-zero specifications. The analysis data were formulated as definitions and rules for the KTH text-to-speech synthesis.

Two perception tests were performed on different versions of the preliminary rule system. In the first test, a subset of the Swedish consonants were tested. The subjects were asked to mark the more natural-sounding of two versions of a consonant in a VCV. The differences between the stimuli pairs consisted mainly of inclusion of the pole/zero pair and modifications of the source parameters. The results were encouraging, even though the inclusion of the pole/zero pair was not always a success. Subjects commented on a high degree of naturalness of the female speech.

The second test was a diagnostic test of the preliminary synthesis-by-rules system. Only a few results are interesting in the context of this paper, for further discussions see [14]. In the new system the voice bars of voiced stops was produced using glottal parameters. This was accepted. This method to produce voice bars will make it easy to vary precision of articulation of stops by voice source variations. The unvoiced stops were sometimes perceived as voiced. This has been corrected by a more abrupt switch-off of the voice source and by adding pre-aspirative noise. The amount of pre-aspiration varies between speakers and is often higher for female speakers. This feature will be used to enhance naturalness.

ACKNOWLEDGEMENTS

This work has been supported by FRN, NUTEK, HSFR, KTH.

REFERENCES

[1] Abadjieva, E., Murray, I. and Arnott, J. (1993): " Applying analysis of human emotional speech to enhance synthetic speech", Proc. Eurospeech '93, pp. 909-912

[2] Cahn, J. (1990): "The generation of affect in synthesised speech", J of the American Voice I/O Society, Vol. 8, pp. 1-19

[3] Carlson R., Granström B., Karlsson, I. (1990): "Experiments with voice modelling in speech synthesis", Speech Communication, Vol. 10, pp. 481-490

[4] Carlson R., Granström B., Nord, L. (1992): "Experiments with emotive speech - acted utterances and synthesized replicas", Proc. of 1992 ICSLP, pp. 671-674

[5] Childers, D. & Lee, C. (1991): "Vocal quality factors: Analysis, synthesis and perception", JASA, Vol. 90, pp. 2394-2410

[6] Fant, G. (1953): Speech communication research. Ingenjörsvetenskapsakademin Stockholm Sweden, 24, 331-337

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

[8] Fujisaki, H. and Ljungkvist, M. (1986): "Proposal and evaluation of models for the glottal source waveform", Proc ICASSP 86, pp. 1605-1608

[9] Holmes, J. (1961) Research on speech synthesis carried out during a visit to the royal institute of technology, Stockholm, from November, 1960, to March, 1961. Res. Rep. Ref. JU11-4, Joint Speech Research Unit, GPO, Eastcote, England

[10] Holmes, J., (1973): "The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer", IEEE Transactions on Audio and Electroacoustics, Vol. AU-21, pp. 298-305

[11] Imaizumi, S., Kiritani, S., and Saito, S. (1991): "Perceptual evaluation of a glottal source model for voice quality control", in Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms, ed.: J Gauffin and B Hammarberg, Singular Publ. Inc., San Diego, pp. 225-233

[12] Karlsson, I., (1991): "Female voices in speech synthesis", Journal of Phonetics, Vol. 19, pp. 111-120.

[13] Karlsson, I., (1992): "Modelling voice variations in female speech synthesis", Speech Communication, Vol. 11, pp. 491-495

[14] Karlsson, I., and Neovius, L. (1994): " VCV-sequences in a preliminary text-to-speech system for female speech", STL-QPSR 1/94, forthcoming

[15] Kasuya, H., and Ando, Y. (1991): "Acoustic analysis, synthesis and perception of breathy voice", in Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms, ed.: J Gauffin and B Hammarberg, Singular Publishing Group Inc., San Diego, pp. 251-258

[16] Klatt, D. (1987): "Review of text-to-speech conversion for English", JASA, Vol. 82, pp. 737-793

[17] Klatt, D., and Klatt, L. (1990): "Analysis, synthesis and perception of voice quality variations among female and male talkers", JASA, Vol. 87, pp. 820-857

[18] Laver, J. (1980) The phonetic description of voice quality. Cambridge University Press, Cambridge

[19] Macchi, M., Altom, M., Kahn, D. Singhal, S. and Spiegel, M. (1993): "Intelligibility as a function of speech coding method for template-based speech synthesis", Proc. Eurospeech '93, pp. 893-896

[20] Pinto, N.B., Childers, D.G. and Lalwani, A.L. (1989): " Formant speech synthesis: improving production quality", IEEE Trans.on Acoustics, Speech, and Signal Processing 37, pp. 1870-1887

[21] Rosenberg, A., (1971): "Effect of glottal pulse shape on the quality of natural vowels", JASA, Vol. 49, pp. 583-590

[22] Rothenberg, M., Carlson, R., Granström, B. and Lindqvist-Gauffin, J., (1974): "A three-parameter voice source for speech synthesis", Proc. Speech Communication Seminar, Stockholm, Aug. 1-3, 1974 , A&W Int., Stockholm, Vol. 2. pp. 235-243

[23] Strevens, P. (1958) VII International Congress of Linguistics, Oslo, August 5-9. Report of Visit by PS and JA, including lecture-demonstration of PAT. In University of Edinburgh, Phonetics Dep., Report on The Specification of Speech Sounds by means of Acoustic Parameters, Edinburgh, Scotland, pp 7-24