Journal of Phonetics (1991) **19,** 111–120

Female voices in speech synthesis

Inger Karlsson

Department of Speech Communication and Music Acoustics, KTH, Fack 70014, S-100 44 Stockholm, Sweden

The intention of this paper is to give a short overview of the history and the present stage of synthesis of female voices. The earlier attempts at synthesizing a female voice consisted mainly of either transforming a male synthetic voice or copying an utterance by a female speaker. These attempts were often not completely successful. This was mainly due to the synthesizers not being complex enough; often only the first three or four formants, fundamental frequency and overall amplitude could be varied. Some of these previous attempts are discussed briefly in the paper.

The more versatile synthesizers available today in combination with new analysis tools have given us the opportunity to synthesize better voices. Data on vowel formants, fundamental frequency and voice source variations for female voices are collected. The differences between female and male speech is discussed briefly in the paper. These data on the female voice are used in the new synthesizers to produce female-sounding synthetic speech. Our current work in this area consists both of copying a female utterance and writing female pronounciation rules for a text-to-speech system.

1. Introduction

The need for voice variations has always been apparent in different speech synthesis applications, such as voice prosthesis and translating telephony. Accordingly, many attempts at synthesizing female voices, child voices and also different voice qualities have been made. So far, these attempts have not been very successful. The reasons for this are both of a technical and a descriptive nature. The earlier synthesizers have not been versatile enough to allow different voice qualities. Nor do we have sufficient data on what consititutes the differences between different voices. However, the more complex synthesizers of today in combination with new software analysis programs are making it feasible to produce many different natural-sounding synthetic voices. In this paper we have concentrated on how female voices have been produced by synthesizers. It also contains a description of some of the work that is currently going on to make a text-to-speech system using a formant synthesizer speak with a voice that is undeniably female.

0095-4470/91/010111 + 10 $03.00/0

2. Historical overview

2.1. *Early attempts*

Synthesis of the human voice by means of electronic devices has a fairly long history. The earlier synthesizers that used only a few parameters to produce the synthesized speech were constructed around 1950, that is the Haskins Laboratories Pattern Playback (Cooper, Liberman & Borst, 1951), PAT (=Parametric Artificial Talker, See Lawrence, 1953) and OVE I (Fant, 1953). The speech that was synthesized with these machines was mostly male-like speech for various reasons: the Pattern Playback could only produce an F_0 of 120 Hz and all three synthesizers mentioned had a fixed higher pole correction that was appropriate for a male voice. Some effort went into producing female-sounding speech, however. In 1956, Fant tried to produce a conversation between a male and a female voice on a tape demonstrating OVE I. When a later version of PAT was presented at a congress of linguistics in Oslo in 1957 (Strevens, 1958), one item of the demonstration was a sentence by a female voice. A later version of OVE, the OVE II synthesizer, was used by J. Holmes (1961) to copy a male and a female utterance. Listeners judged the male copy to be much closer to the copied natural utterance than the female copy. This could depend on the choice of speech material but a more likely reason is the constraints built into the synthesizer. Listening to these early trials of female voice synthesis today one realizes that none of them sounded convincingly female.

2.2. *Copy synthesis*

In all these early examples a fairly crude model of the voice source and often a fixed setting for formants above the third formants was used, which is a probable explanation for the failures. With the development of more sophisticated synthesizers, where particularly the voice source has become much more versatile, the synthesis of more than a single male voice is becoming feasible. In recent years some promising results have been achieved in copying natural female speech. Klatt (1986) has produced a natural sounding copy of some vowels and one sentence of reiterant speech uttered by a female speaker using a source where some spectral properties and the noise content of the voiced source was varied. Fant, Gobl, Karlsson & Lin (1987) copied some sentences by female speakers using the LF-model of the glottal source (Fant, Liljencrants & Lin, 1985), and achieved good naturalness. W. J. Holmes (1989) has experimented with copying female speech with a parallel formant synthesizer which also produced female speech of acceptable quality. Recently we have used a extended version of the OVE III synthesizer, called GLOVE, including the LF-model voice source, to successfully approximate a female sentence (Carlson, Granström & Karlsson, 1990). This last experiment is described more fully below. These are only some examples of the work on copy synthesis that is going on.

2.3. *Synthesis by rule*

An overwhelming part of the work invested in producing synthetic voices for text-to-speech synthesis have so far been concentrated on male voices. Even today, the so far most successful synthetic female voice, the female voice of the DECtalk is a simple transformation of the original male voice of the system (Klatt, 1987), and

not built on analysis of a natural female speaker. Klatt's male-to-female transformation consisted in scaling F_0 by a factor 1.7 and the formants by 1.175. He also removed the fifth formant and increased the open quotient; both these things would alter the spectral tilt. According to Klatt "These manipulations are not sufficient to turn [the male voice] into a convincing female speaker" (Klatt, 1987, p. 784). Similar transformations are included in the multilingual INFOVOX text-to-speech system based on previous work at our department.

2.4. *Voice transformation*

Some researchers have tried to transform a male into a female voice or the opposite starting from an LPC coded version of conversational speech. Traunmüller, Branderud & Bigestans (1989) transformed only pole values and fundamental frequency, using a transformation formula that raised lower frequencies more than higher frequencies in a conversion from a male to a female voice. Amplitudes, durations, and so on were kept at the original values. Such procedures will sometimes give astonishingly good transformations, especially when going from a female to a male voice. The perceptual shift between male and female voices was helped by some manipulation of the fundamental frequency contour. Childers and his co-workers have studied voice transformation. They have transformed the shape of the voice pulse together with formants, bandwidths and fundamental frequency using formant synthesis (Childers, Wu, Hicks & Yegnanarayana, 1989; Pinto, Childers & Lalwani, 1989). They claim that it is important to get a good transformation of the voice pulse shape as well as of formants and fundamental frequency to achieve an acceptable transformation. A copy of the intended speaker's fundamental frequency contour will improve the transformation. They also found it easier to get a convincing transformation going from a female to a male voice than in the opposite direction.

3. Voice quality variations

The above examples of female voice synthesis have all concentrated on the differences in fundamental frequency, formant frequencies and the spectral tilt of the harmonics. There might also be other differences in quality between male and female voices, which can have physiological or sociological origins. Klatt & Klatt (1990) have found for American English speakers that the average female speaker has more breathiness in her voice production than the average male speaker. This is manifested in a higher degree of noise excitation of the higher formants and of less harmonic energy in these formants. They have tried to synthesize this breathy voice quality difference in reiterant speech by varying different voice source parameters: noise content, spectral tilt, amplitude of first harmonic. In a listening test where the listeners judged the naturalness of the utterance the results indicate that the noise content and the spectral tilt were important to add naturalness to an utterance synthesized using female formant and F_0 values.

4. Acoustics of the female voice

Some of the reasons for the difficulties with synthesis of female voices, can be found in the acoustics of speech. The higher fundamental frequency for a female voice

results in a greater uncertainty in the specification of formant values. The average F_0 difference between men and women is about 0.9 octave, while the formant bandwidths are only about 20% higher for women. This implies that fewer harmonics will fall within the formant bandwidth for female than for male speakers, and accordingly the formants will be less well specified. For high vowels and voiced consonants uttered by women the first formant is often very close to the fundamental of the voice source spectrum. This makes it harder to measure the first formant accurately and also to separate source and vocal tract acoustically. Another difficulty is that the range of possible voices for females is restricted at one extreme by male voices, and at the other extreme by child voices. This implies that listeners will be more critical towards a female synthetic voice than towards a male or even a child synthetic voice.

Over the years, data on male–female voice differences have been collected. Vowel formant frequency and fundamental frequency differences have been studied extensively. Classic work in this area includes the Peterson & Barney (1952) study of vowels uttered by male, female and child speakers, and the Fant (1959) study of male and female vowels. The male–female differences in vowel formant frequencies depending on place and manner of articulation for different languages have been investigated by Fant (1975). He found that the female formant values were on the average 17% higher than the male values, but also that the difference varied with articulation. High back rounded vowels, for example, showed much smaller differences for both the first and the second formant, while open unrounded vowels showed larger differences for the first formant.

Voice source differences have been studied using inverse filtering. Monsen & Engebretson (1977) found that the voice pulses of women were more symmetrical and contained less high frequency components than those of men. As their female informants spoke with considerably weaker voices than their male informants, it is hard to know how much the differences are due to physiology and how much to learned behavior. A clear difference in high frequency content of the voice source for female and male speakers has also been found by Price (1989) studying American English speakers uttering vowels. In a paper by Gobl & Karlsson (1989) concerning Swedish speakers pronouncing sentences, the difference in high frequency content was not found to be very great.

The average fundamental frequency differences between male and female speakers vary with language (culture) but seem often to be close to 90% of an octave higher for women. For a long time there has been controversy over whether frequency span in normal speech is larger for females; Henton (1989) gives a review of this discussion. Data suggest that if the fundamental frequency is measured in semitones, the span is equal for the average male and female speaker, even though individual speakers differ (see for example Fant & Kruckenberg, 1989, and Tielen, 1989). There might be differences in the use of fundamental frequency variations between men and women (see Brend, 1971); these types of differences are culturally dependent.

There are also other culture-dependent differences between the speech of women and men; for example it has been found that women tend to conform more to standard pronunciations and men to talk with a broader dialect, and that women tend to have weaker voices. There are also differences in how the language is used, these aspects fall outside the theme of this paper. For a discussion of them see Thorne & Henley (1975).

5. Present work

The work we are currently involved in is foremost to write rules and decide parameter values of segments for a normal female synthetic voice for a text-to-speech system. The system is built on our text-to-speech program and an extended version of OVE III called GLOVE. GLOVE contains the LF-model voice source and facilitates modulation of the noise source, mixing noise with voiced excitation, a more dynamic variation of the higher pole correction and also two pole/zero pairs in nasal sounds. The first formant bandwidth can be varied with glottal opening, and the rate of change of formant transitions can be regulated with greater accuracy than in former versions of our text-to-speech system. For a detailed description of the LF model see Fant *et al.* (1985), and of GLOVE see Carlson *et al.* (1990). This system has also been utilized to approximate by synthesis a sentence pronounced by a female speaker. This was done to get some ideas of the feasibility for using the current synthesis system for female voice production.

5.1. *Synthesis by analysis*

The natural Swedish sentence /ˇpiːa ˇuːdlar ‹bloː vɪ‹uːler/ uttered by a female speaker was selected for copying. The speaker had what was judged by a speech therapist to be "a dark, slightly coarse, sonorous voice, and swollen vocal cords", and was found in a recording of /pa/-syllables with the Rothenberg mask to have a large amount of residual flow in the closed portion of the voice pulse (Karlsson, 1988). This indicates that the vocal cords are never completely closed and that a coupling to the subglottal system can be expected. Accordingly, there will be extra pole/zero pairs in all voiced segments.

The sentence was analysed in terms of segment durations, formant frequencies, fundamental frequency and glottal source parameters. Linear interpolation was used to define the formant movements between extreme values. The sentence was inverse filtered at a few positions in each phoneme, typically in the middle of a phoneme, at the boundary between two phonemes, and in the most constricted part of the long /u/, /o/ and /i/ vowels. The inverse-filtered pulse was matched by a voice pulse defined by the LF-model parameters to get the voice parameters for the synthetic source. The measured specifications were used as control parameters for the GLOVE synthesizer, and a synthetic representation was produced. It was found that the quality was not a good enough female voice quality.

Several aspects had to be modelled in addition to the formant values and the voice source parameters. The new higher pole correction proved to be a valuable improvement for the spectral balance. The default bandwidth of the first formant was too narrow. (Without adjusting this value the synthesis was perceived as speech in a closed room or resonator.) Letting the first formant bandwidth vary with glottal opening improved the quality. This has approximately the same effect as broadening the bandwidth (Nord, Ananthapadmanabha & Fant, 1986).

The major difference between our speaker and the synthesis was the leaky voice source of the speaker. The noise and the pole–zero pairs due to the glottis never closing completely had been a problem already in the inverse filtering phase. When noise was added pitch synchronously, the similarity between the synthetic copy and the original was enhanced. The addition of noise was specially important to mark the juncture between the /iː/ and the /a/ and in the final part of the sentence.

Addition of noise in the vocalic consonants in some contexts was also of importance as in the first two examples of /l/ in the spectrogram in Fig. 1. However, this type of noise is probably of a different nature and can often occur in voices with very little glottal leakage in other contexts. An /l/ sound after a stop usually contains frication, but if a morph boundary is intervening, the /l/ often lacks frication.

After including these new parameters in the synthesis, the copy was judged to be perceptually very similar to the original in an informal listening test. The remaining differences are very much due to the sketchy copying of the voice source. Spectrograms of original and synthetic speech are found in Fig. 1.

5.2. *Text-to-speech synthesis*

The work on rules for our text-to-speech system has so far been concentrated on definitions of voice source and vocal tract parameters. We have chosen to use

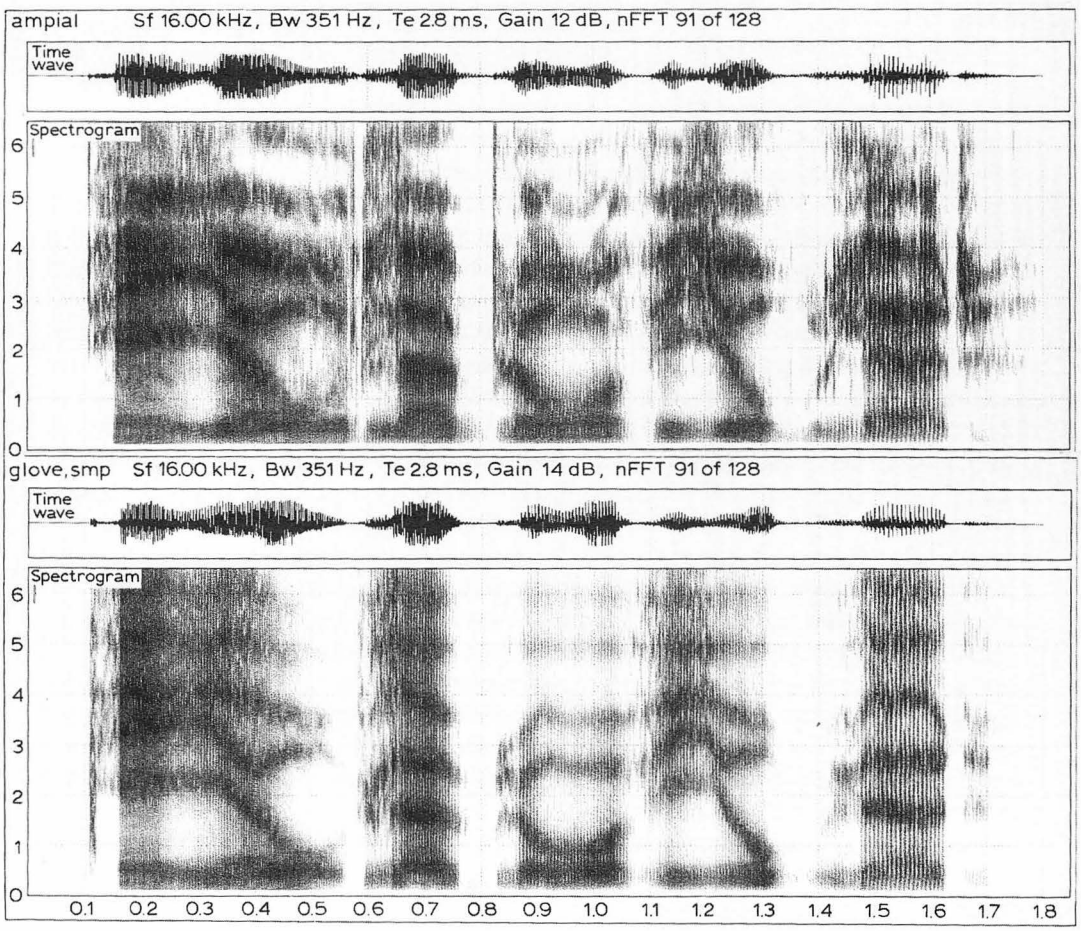

Figure 1. Time wave and broad band spectrograms of the Swedish sentence /"pi:a "u:dlar 'blo: vɪ'u:ler/ uttered by a female speaker (top) and produced by the extended synthesizer GLOVE (bottom).

parameter values measured from the speech of one specific female speaker rather than trying to do transformations of the existing male synthetic voice. This speaker has been judged by a speech therapist to have "a normal, somewhat tight, sonorous" voice (Karlsson, 1988). Her vowel formants were found to be on the average 14% higher than for the male synthetic voice, which is slightly lower than the average male–female difference. Her fundamental frequency is 0.9 octave higher than the fundamental frequency of the male synthetic voice, which is a normal male–female difference. This speaker showed little or no noise content in vowels, contrary to the female speakers of Klatt & Klatt (1990). A comparison between the excitation of the fourth formant in a vowel uttered by this female voice and a female voice containing a large amount of noise is shown in Fig. 2.

At the present stage we presume that there are only small, even negligible, differences between a normal male and a normal female speaker concerning phoneme durations, degree of reduction and coarticulation, stress placement, etc. Accordingly, most of the rules governing duration, degree of coarticulation and overall fundamental frequency movements are the same as for the male voice, even though the absolute values for formants and fundamental frequency are different. In a study of readings of a short text by 11 men and 4 women, Fant & Kruckenberg (1989) found no sex related variations in fundamental frequency range measured in semitones, in phoneme duration, or in pause length. The differences in fundamental frequency contours that improved the voice transformations discussed above have so far not been tested in the synthesis. It was found, though, that the "femaleness" of the speech was slightly enhanced if F_0 was raised for the very last pitch periods of a sentence.

For the different speech segments, formant frequencies have been measured both from ordinary broad band spectrograms and by matching spectra from natural speech with synthetic speech. The matching procedure was particularly useful for achieving the formant bandwidths both in voiced and unvoiced segments, and also for deciding the source amplitudes for the unvoiced sounds. Voiced segments were inverse filtered and the inverse-filtered voice pulses were matched by a LF-model source pulse. Thus voice source parameter values were collected. The inverse filtering also gave formant frequencies, bandwidths and fundamental frequency.

On the phrase level, data on voice source variations due to coarticulation, stress and termination of the phrase have been collected. From the same sentence material, data on sentence intonation and on F_0 movements and range in focally

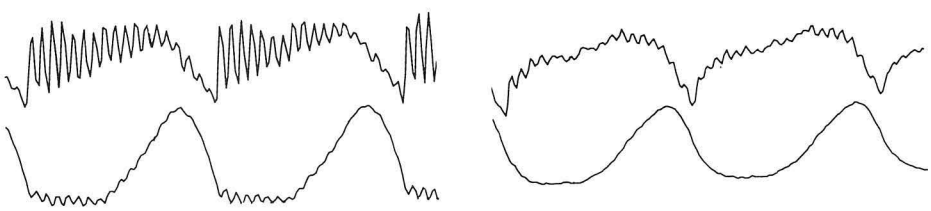

Figure 2. Examples of voiced (left) and noise (right) excitation of the fourth formant. The examples were taken from the same vowel in the same context uttered by two female speakers. The top curve represents the pressure wave with all formants but the fourth cancelled. The bottom curve is the integral of the top curve, which is equal to a high-passed (above 10 Hz) representation of the glottal flow with only the fourth formant excitation remaining.

stressed words have been determined. Some results from these measurements both for isolated syllables and connected speech are discussed in Karlsson (1989).

5.2.1. *Evaluation*

Data on both segments and phrases are included in a set of definitions and rules for a female voice for our text-to-speech system. The different phonemes are then synthesized. The vowels are synthesized one by one in isolation. The consonants are put in a V–V context. Informal listening tests are made, keeping in mind that the voice should sound female and that at the same time the intended phoneme should be perceived. The segments are also compared to natural speech with the help of spectral sections. If necessary, the parameters are adjusted to get a better fit between natural and synthetic sounds. Care is taken that the voice really sounds female. So far, the voiced vowel sounds do not contain noise in the higher frequency range, which according to Klatt & Klatt (1990) might be an important cue for female voice quality. The main reason for this is that it does not seem necessary to use noise excitation in vowels to achieve a female sounding synthesis modelled on the particular female prototype we have chosen. We plan to test the importance of noise excitation of higher formants in the near future, though.

The resulting formant frequencies for the vowels are plotted together with the formant frequencies for the male synthetic voice in Fig. 3. The differences between the male and the female voice comply very well with the differences found by Fant (1975) comparing female and male speakers of different languages. He found that for open vowels there is a large difference in the first formant, that for high, front vowels there is a large difference in the second formant, while for the back rounded vowels all differences are small. To achieve vowels that were perceptually convincing, however, a change of formant values was not enough. The default bandwidth of the first formant was often too narrow, especially in the more open

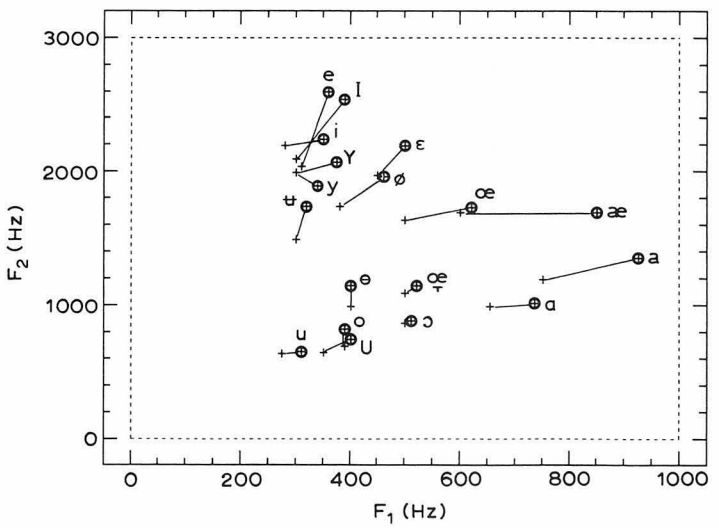

Figure 3. First and second formants for the female and male synthetic vowels. The female values are marked by a ring. The vowel pairs are joined by a line and the IPA symbol is indicated next to the female values. (For the front vowels, the third formant carries much information.)

vowels. The voice source also turned out to be different for different vowels, the return time parameter of the LF-model that influence the spectral tilt was smaller for more open vowels (Karlsson, 1989). The voice pulse for open vowels should consequently contain more high frequency energy than for closed vowels. These differences were of perceptual importance. Presently, the work is concentrated on consonants; we have so far no results to report on these.

The overall sentence intonation pattern that is used for the male voice could be transformed to a female voice. The average fundamental frequency was raised by a factor of 1.9, and the frequency range in octaves was kept. The only difference in overall pattern that has been implemented so far is that the final fall of fundamental frequency is interrupted close to the end of phonation and the fundamental frequency is raised slightly.

6. Conclusions

For as long as researchers have tried to produce human speech by machine they have also tried to make the machine speak with different voices. For a long time the female voices did not really sound female. One reason for that was that the machines were primarily constructed to produce male-like speech (his master's voice). Accordingly, some restrictions that were built in made the machines unsuitable for female voice production. Also, most analysis of speech was done on male speech. The lack of data on female speech caused additional obstacles. As more and more sophisticated speech synthesizers and analysis tools become available, the female synthetic voices are sounding more female. Today, we are able to produce a very good female voice by synthesis by analysis but there is still some work to be done before we have fully achieved our goal: a text-to-speech system using a formant synthesizer that can speak with different voices, male, female and child, and with varying voice qualities within these major groups. With the new, extended synthesizer it seems to be possible to achieve this; what we lack now is mainly data to describe the different voices and voice qualities.

The female voice synthesis project has been supported in part by grants from the Swedish Board for Technical Development (STU) and Swedish Telecom.

References

Brend, R. (1971) Male–female intonation patterns in American English. In *Proceedings of the seventh international congress of phonetic sciences*, pp. 866–869 The Hague: Mouton.

Carlson, R., Granström, B. & Karlsson, I. (1990) Experiments with voice modelling in speech synthesis. In *Proceedings of the tutorial and research workshop on speaker characterization in speech technology*, Edinburgh 26–28 June 1990, pp. 28–39.

Childers, D. G., Wu, K., Hicks, D. M. & Yegnanarayana, B. (1989) Voice conversion, *Speech communication*, **8**, 147–158.

Cooper, F. S., Liberman, A. M. & Borst, J. M. (1951) The interconversion of audible and visible patterns as a basis for research in the perception of speech, *Proceedings of the National Academy of Science (U.S.)*, **37**, 318–325.

Fant, G. (1953) Speech communication research, *Ingenjörsvetenskapsakademin Stockholm Sweden*, **24**, 331–337.

Fant, G. (1975) Non-uniform vowel normalization, *Speech transmission laboratory, quarterly progress and status report (KTH, Stockholm)* No. 2–3, 1–19.

Fant, G. Liljencrants, J. & Lin, Q. (1985) A four-parameter model of glottal flow, *Speech transmission laboratory, quarterly progress and status report (KTH, Stockholm)* No. 4, 1–13.

Fant, G., Gobl, C., Karlsson, I. & Lin, Q. (1987) The female voice–experiments and overview, *Journal of the Acoustical Society of America,* **82** (Suppl. 1), S90.

Fant, G. & Kruckenberg, A. (1989) Preliminaries to the study of Swedish prose reading and reading style, *Speech transmission laboratory, quarterly progress and status report* (*KTH, Stockholm*) No. 2, 1–80.

Gobl, C. & Karlsson, I. (1989) Male and female voice source dynamics. In *Proceedings of vocal fold physiology conference, Stockholm* (to appear).

Henton, C. (1989) Fact and fiction in the description of female and male pitch, *Language and Communication,* **9**, 299–311.

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. Research report reference JU11-4, Joint Speech Research Unit, General Post Office, Eastcote, U.K.

Holmes, W. J. (1989) Copy synthesis of female speech using the JSRU parallel formant synthesizer. In *Proceedings of the European conference on speech communication and technology,* Paris 1989, Vol. 2, pp. 513–516.

Karlsson, I. (1988) Glottal waveform parameters for different speaker types. In *Proceedings of Speech '88, 7th FASE symposium,* Vol. 1, pp. 225–231.

Karlsson, I. (1989) A female voice for a text-to-speech system. In *Proceedings of the European Conference on Speech Communication and Technology,* Vol. 1, Paris 1989, 349–352.

Klatt, D. (1986) Detailed spectral analysis of a female voice, *Journal of the Acoustical Society of America,* **81** (Suppl. 1), S97.

Klatt, D. (1987) Review of text-to-speech conversion for English, *Journal of the Acoustical Society of America,* **82**, 737–793.

Klatt, D. & Klatt, L. (1990) Analysis, synthesis and perception of voice quality variations among female and male talkers, *Journal of the Acoustical Society of America,* **87**, 820–857.

Lawrence, W. (1953) The synthesis of speech from signals which have a low information rate. In *Communication theory* (W. Jackson, editor), pp. 460–469. London: Butterworths.

Monsen, R. & Engebretson, A. (1977) Study of variations in the male and female glottal wave, *Journal of the Acoustical Society of America,* **62**, 981–993.

Nord, L. Ananthapadmanabha, T. & Fant, G. (1986) Perceptual tests using an interactive source filter model and considerations for synthesis strategies, *Journal of Phonetics,* **14**, 401–404.

Peterson, G. & Barney, H. (1952) Control methods used in a study of the vowels, *Journal of the Acoustical Society of America,* **24**, 175–184.

Pinto, N. B., Childers, D. G. & Lalwani, A. L. (1989) Formant speech synthesis: improving production quality. *IEEE Transactions on acoustics, speech, and signal processing,* **37**, 1870–1887.

Price, P. (1989) Male and female voice source characteristics: Inverse filtering results, *Speech Communication,* **8**, 261–277.

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

Thorne, B. & Henley, N. (1975) *Language and sex: difference and dominance,* Rowley, MA: Newbury House.

Tielen, M. (1989) Fundamental frequency characteristics of middle aged men and women. In *13th proceedings of the Institute of Phonetic Sciences, University of Amsterdam,* pp. 45–58.

Traunmüller, H., Branderud, P. & Bigestans, A. (1989) Paralinguistic speech signal transformations, *Phonetic Experimental Research at the Institute of Linguistics University of Stockholm,* **10**, 47–64.