SEX AND SPEECH SYNTHESIS: TECHNIQUES, SUCCESSES AND CHALLENGES

Caroline Henton

Advanced Technology Group
Apple Computer Inc.

ABSTRACT - Female speech synthesis has a short history. Its quality is marginal in most current systems. Examples of synthetic speech will be played, indicating remaining problems in synthesizing female speech. A template for female speech is given, together with a review of applications for synthetic speech.

INTRODUCTION

Speech synthesis has evolved from manually-intensive techniques to current sophisticated synthesis-by-rule, and simpler 'paint by numbers' concatenative systems. After 50 years of such work, we expect to see signs of relative maturity. In no current system, using any of the varied techniques - parametric, concatenative, articulatory modelling, hybrid - does female speech sound natural or bearable for utterances any longer than brief comments. Few technological and signal processing challenges remain in generating tolerable female speech. So why is synthetic female speech so unconvincing?

This paper surveys the representation and successfulness of female synthesis. Phonetic ingredients are used to draw a profile of successful female synthetic speech. Segmental synthesis is not discussed, since that challenge has been largely met: intelligibility of single words and segments has reached asymptote, lying at a reliable 95% level for some years. Goals for speech synthesis lie elsewhere: in achieving natural voice quality, pleasantness, avoiding monotony, and in characterizing different speaking styles. Applications for female speech are also listed, and suggestions are made about the (in)appropriateness of sex-specific synthetic or natural speech.

FEMALES IN SPEECH SYSTEMS

A Hypercard stack containing 18 examples of synthesizers will be played. Only 6 systems offer a female voice as an alternative to the male, or exclusively. It is perhaps indicative of the continuing lack of acoustic knowledge about female speech that only two of the systems are parallel-formant synthesizers, the others all use real female speech in concatenative diphone systems. From this it might be inferred that it is easier to chop up speech and re-combine waveforms than it is to determine acoustic parameters needed to produce female speech by rule.

A FEMININE ABSENCE

The paucity of female synthetic speech is attributable to two sources. Firstly, females have been the 'second' or missing sex in scientific research in most disciplines. Secondly, speech research has androcentric foundations. Minimal data has been assembled for the speech of 51% of adult speakers (Henton, 1987). Recent data collection efforts have tried to equalize numbers according to sex, and many U.S. research agencies now require 'gender equality' in experimental populations. Nevertheless, in acoustic phonetic research, the female voice has been excluded or marginalized for decades. The mechanical origin of the 'problem' lay with the primary tool for speech analysis, designed by men for the analysis of male speech (Klatt, 1982, p.83). Analysis bandwidths of spectrographs are now minutely adjustable, and other computational means of digital signal processing (e.g. LPC) exist for exploring female speech, but the residue of resistance and the notion that there is something 'more difficult' (not just different) about female speech persist.

Databases should demand equal numbers of the sexes, should be larger, and more representative of females' accents, SES, and geographical origin. Profiling female voice and/or speech requires more than piecemeal, haphazard observations of one speaker. Attempts to model the archetypal female speaker, give rise to other problems. While speech therapists use protocols that distinguish normal from abnormal voices, and phoneticians and lay listeners alike can express opinions and preferences about the 'acceptability' of a voice, it seems there has been no determination of an 'optimal' voice quality that might be used in speech synthesis. There are no studies in which the parameters of parallel-formant synthesizers are manipulated to produce a continuum of voice qualities ranging from

the minimally acceptable to the maximally desirable voice. These and other studies have to be conducted before researchers can determine what it is they are trying to capture in a pleasant, synthetic voice. From these findings it will also be possible to provide 'customized' voice qualities according to the needs and predilections of different users. Research into female speech has appeared in the past decade, yielding important analyses, but the 'lag time' in incorporating this knowledge into commercial synthesizers is frustratingly slow. Some salient acoustic parameters and prosodic patterns that contribute to an optimal female voice are given below.

Invariance of articulatory 'space' and in articulatory precision

Formant frequencies undoubtedly carry less perceptual weight in distinguishing speakers than do phonation differences (cf. Coleman, 1976). The overall articulatory space created by formant frequencies may nonetheless be important in a sex-specific profile. Across 7 languages and dialects, females' vowels (after normalization) are both closer to the periphery of the vowel space, and comparatively 'wider' in the F1 dimension (Henton, 1992a). Females produce more 'open-mouthed' variants of peripheral vowels - behavior that may be equated with greater articulatory distinction, careful pronunciation, or intelligibility. With further data we might deduce whether in synthesis we need female F1 parameters that are greater than might be expected, in comparison to values in a male voice.

Voice quality, voicing, and their distributions within utterances

Until recently, parallel-formant synthesizers produced male voices alone. It then occurred to researchers that females might want 'a voice of their own'. One approach took parameters for male speech, and 'adjusted' them to produce a sex change. This approach, as explored by Barry (1986), Klatt (1986) Pickering (1988) and Karlsson (1986-92), of 'model by adjustment' was profoundly flawed. The 'Frankinstein' approach - making a woman's voice out of a man's - created mutant androgyny rather than a 'warm' female (cf. Klatt's pessimistic conclusion, 1986). Simply doubling the frequency of a male voice SFF and manipulating the formant frequencies will not engender a female voice. Speakers show considerable variation according to gender, language, accent, SES, age, health, and degree of socially-induced behaviour. The creation of a synthetic voice must take into account the gender boundaries of the society in which the voice will be used.

Regarding naturalness, the shape of the glottal excitation wave contributes more to speech quality (Childers and Wu, 1990) than simple changes to other parameters. Studies of females' and males' glottal gestures have revealed marked differences (Cheng and Guerin, 1987; Holmberg, et al., 1988; Javkin et al., 1989; Karlsson, 1986, 1989; Klatt, 1986; Pickering, 1988; Titze, 1988). Female speakers' waveforms are more symmetrical, with no peak during the opening phase. Females' glottal waveforms have gentler closing phases (20-40% of the total period), with a more rounded 'shoulder' at the end of the phase. Along with lower vocal fold closing velocity and lower ac flow, normal and loud female voices have a shorter closed phase, which may result in a steeper spectral slope. In many females' waveforms, there may be no definite closed period at all.

In addition to the engineering precision needed to model glottal movements, there are questions about where, and how frequently, voice quality effects appear in utterances. Baseline information about the sociophonetic production and sex-specific use of voice qualities in extended discourse is slow to emerge. In studies of American English speakers, the voice quality judged to be characteristic of female voices was nasality; for males, hoarseness (Murry & Singh, 1980). But different voice qualities may be gender-linked differently across cultures. Henton & Bladon (1985; 1986) examined the speech of 80 British English speakers for breathiness, and creak. Females used significantly greater amounts of breathiness (in open vowels). In American English, females are also breathy (Klatt, 1987; Klatt & Klatt, 1990). In their rule-based female voices, Karlsson (1989) and Klatt & Klatt (1990) control breathiness using multiple parameter glottal sources. When a successful breathy voice is available, its intensity and frequency of occurrence in utterances should, however, be variable (cf. Karlsson, 1992). An excessively breathy voice or its wholesale use in discourse may be perceived as unpleasant, even pathological.

Creak was also sex-marked in British English, with males using dramatically more creak. Henton & Bladon (1986) found creak occurred most frequently utterance-finally, carrying the linguistic function of pre-pausal demarcation; and it was employed to different degrees in different accents, possibly to indicate hypo/hyper-masculinity. These findings were also borne out for male speakers of American English (Henton, 1989b), for whom creak was almost compulsory in V-V transitions across word

boundaries, and, in some speaker styles, occurs in varying degrees throughout the speaker's speech, appearing to be a 'vocal setting'. Informal observation of Australian English further supports this use of creak as a marker of masculinity. Cross-language transfer must not however be expected. For 4 speakers of Swedish, Huber (1988) found significant sex-specific use of creak, but there the females used more diplophonic phonation and creak, whereas the males preferred creaky voice utterance- finally. Naturalness in synthesis will benefit from judicious use of creak; female synthesis might benefit most from limiting its use to utterance-final position.

We clearly need more research into the socio-cultural and pragmatic use of voice qualities. In addition to the voice qualities mentioned above, examination of pharyngealization, rhoticity, and whisper might be productive. It may be further hypothesized that comparative intensity, and dispersal, of voicing, devoicing, and voicelessness through extended utterances will vary according to gender. Strength of voicing is thus a further avenue for exploration.

Differentiating prosodic features

Speaking fundamental frequency, pitch range and dynamism, voice quality, phrasing, rate, hesitancy, intensity, and intonational tunes are all known to be used as clues in differentiating voices (see Henton, 1987; McConnell-Ginet, 1983, for overviews). Despite a welter of sex-specific results, workers in speech synthesis have been disappointingly slow to acknowledge that separate prosodic settings are needed for female voices. Crystal (1975) notes anecdotally that female/male speech differs in glissando effects, complex-tone usage, breathiness, and moving to falsetto. Experimental results support some of these observations. Huber's two female speakers of Swedish used "more intonation tunes per text, involving higher onsets and offsets, shorter durations, steeper falls, a larger proportion of rising versus falling declination lines..." Tielen (1992) reports that female speakers of Standard Dutch speak at a faster rate than do males, all other things being equal. Intensity means and ranges in females' speech are less than males' (Markel et al., 1972; Günzberger, 1987), in accord with feminine stereotypes of "soft and low" voices. Females generally show greater verbal fluency than males (see review in Henton, 1992b). Pitch range for both sexes is essentially similar (Henton, 1989), but females may be more dynamic within that range (McConnell-Ginet, 1983). Modal values of FO also reveal differences, with females spending more time proportionally in the upper part of their F0 range. Word and sentence stress have been called exaggerated in the speech of females (Jespersen, 1922; Lakoff, 1975), but these intuitions have not been examined in 'normal' speech. Isolated word durations may be longer for females (Günzberger, 1987), possibly correlated with hyper-articulation.

For British English, Pellowe & Jones (1978) and Woods (1992) reported similar findings to Huber's, viz. women use greater intonational variety and more rising tunes than men; further that rising tunes are not linked to syntactic interrogatives. Woods (1992) records women as using more fall-rise and high fall tones than men, who, conversely, use significantly more level tones than do women. Frequent level tones correlate with monotonicity, a pattern found to be a characteristic of males' speech (Bennett and Weinberg, 1979). Too generous monotonicity in synthetic speech, though, can only strengthen the often-heard complaint that a voice is mechanical or robotic. The female habit of changing pitch and loudness frequently has communicative importance in attracting and keeping listeners' attention. Given the low boredom threshold reported for synthetic speech, it is vital that such prosodic variability be included in the synthetic speech of females and males alike. Table 1 provides indications for better archetypal synthetic female (English) speech; factors are rank-ordered; comparisons are to male speech.

TABLE 1. Template for female speech

- Voice quality
  - breathy, with varying amounts for discourse dynamics
  - limited use of creak
  - no excessive nasality, tenseness, or sonority
  - glottal waveform with
    - greater symmetricality
    - less steep closing phase
    - geater open quotient
    - corner-rounding
    - steeper spectral slope
    - greater variability
- Pitch
  - SFF at about 220Hz.
  - pitch range about 0.8 octave

	- greater dynamism of F0
	- greater part of time in upper part of pitch range
• Prosody	- greater number and variety of intonational tunes per text
	- higher onsets and offsets
	- more fall-rises (BE) ; rises (AE)
	- steeper falls
	- faster tempo (but longer vowel durations?)
• General Acoustics	- higher formant frequencies than expected from normalizing from male values
	- wider formant bandwidths
	- lower average intensity
	- greater intensity variation and rate of change
	- less consistent and less constant voicing
	- more peripheral vowel articulations
	- fewer weak forms and/or reduced vowels } more careful
	-greater aspiration and more released stops } pronunciation

APPLICATIONS FOR SYNTHETIC SPEECH

Synthetic speech applications are growing rapidly. Table 2 lists some current and projected ones. Intended use in an application may influence the choice of synthesis technique. Currently, parallel formant synthesis is the only approach that allows users control over different voice qualities, speech style, and minute acoustic adjustments; the naturalness of the voice source may nonetheless remain unsatisfactory for some time. Foreign languages can also, theoretically, be produced more efficiently using this technique. Concatenative synthesis is appropriate if the application must run in software only and if the output is shorter and more stylized than the extended speech needed by some users.

TABLE 2. Uses for synthetic speech

		Currently using female speech?
Aids for:	verbally-impaired	√
	vision-impaired	√
	hearing-impaired	√
Warning devices:	military	?
	medical	?
	heavy industry	?
Advisory messages:	cars	√
	appliances	√
	public transport - airports	√
	- buses	√
	- trains	√
Commercial:	text-to-speech - alerts	?
	- help	√
	- proofreading	√
	- remote information access	√
	financial transactions	√
	inventory tracking/ bar-code reading	√
	telephone directory assistance	
	- numbers	√
	- reverse directory	?
	- zip codes	?
	- teleconferencing	
	toys	√
Educational:	early learning	√
	ESL and foreign language learning	?
	adult literacy	?
	books on disc	
	interactive assignments	
	dialogue creation	
	phonetics, speech pathology	

By the year 2000, the role of speech will have probably expanded into many aspects of daily life. Some predictions are worth considering. Domestic appliances (microwave, dishwasher, lights, garage door) and entertainment devices (stereo, VCR, TV) will use universal voice input controls and synthetic speech to confirm actions . Cars will include speech as a safety option (a 'driver's side windbag'?!), to eliminate many routine tasks that currently require the driver to lift hands from the steering wheel and eyes from the road: using cellular phones, faxes, controlling the sound system, windows, mirrors, sunroof, etc. Synthetic speech may provide guided navigation and a way of interacting with a 'head-up' display. Portable computers are likely to be keyboard-free, receiving and exchanging information by voice and pen. Telephone calls will be placed and answered by the computer with a combination of speech recognition and synthesis. Placing orders at drive-up restaurants and banks will also involve speech recogniton, and confirmation by synthetic voice. Digitized real speech would be better used in some scenarios, owing to the the restricted number and content of the messages. Nevertheless, with this 'future' only 8 years away, it is essential that natural, appealing and appropriate synthetic speech stands a chance of passing the 'Turing test', and that users are *en*abled by synthetic speech, rather than wishing to *dis*able it as soon as possible.

REFERENCES

Barry, M.C. (1986) "Synthesising female voice quality", Cambs. Papers in Phonetics and Experimental Linguistics, 5, 1-14.

Bennett, S. & Weinberg, B. (1979) "Sexual characteristics of pre-adolescent children's voices," Jour. Acoust. Soc. Am., 65: 179-89.

Cheng, Y.M. & Guerin, B. (1987) "Control parameters in male and female glottal sources," in T. Bear, C. Sasaki & K. Harris (eds.) *Laryngeal Function in Phonation and Respiration*, pp. 219-238, (College Hill, San Diego).

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

Coleman, R.O. (1976) "A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice", Jour. Sp. Hear. Res., 19, 168-180.

Crystal, D. (1975) *The English Tone of Voice*, pp.85-86, (Arnold: London).

Drullman, R. and Collier, R. (1991) On the combined use of accented and unaccented diphones in speech synthesis. Jour. Acoust. Soc. Am., 90,1766-1775.

Günzberger, D. (1987) "Duality in vocal gender roles", Prog. Rep. Inst. Phonetics, Utrecht, 12, 1-10.

Henton, C. (1992a) "Acoustic variability in the vowels of female and male speakers", Jour. Acoust. Soc. Am., 91 (4,2), 2387 .

Henton, C. (1992b) "The abnormality of male speech", pp.27-59, in G. Wolf (ed.) *New Departures in Linguistics*, (Garland Publishing: New York).

Henton, C.G. (1987) Phonetic considerations for the synthesis of female voices. *Procs XIth. Intl. Cong. Phonetic Sciences.* Tallinn, Estonia, pp. 270-273.

Henton, C.G. (1989a) "Fact and fiction in the description of female and male pitch", Lang. and Comm., 9, 299-311.

Henton, C.G. (1989b) "Sociophonetic aspects of creaky voice", Jour. Acoust. Soc. Am., 86, S26.

Henton, C. G. & Bladon, R.A.W. (1985) "Breathiness in normal female speech: inefficiency versus desirability", Lang. and Comm., 5, 221-227.

Henton, C. G. & Bladon, R.A.W. (1986) "Creak as a sociophonetic marker", pp. 3-29, in L. Hyman & C.N. Li (eds.) *Language, Speech and Mind: Studies in Honour of Victoria A. Fromkin*, (Routledge: London).

Holmberg, E.B., Hillman, R.E. & Perkell, J.S. (1988), "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice", Jour. Acoust. Soc. Am. 84, 511-529.

Huber, D. (1989) "Voice characteristics of female speech and their representation in computer speech synthesis and recognition', Procs. Eurospeech '89, Paris, 477-480.

Javkin, H. et al. (1989) A multi-lingual text-to-speech system. Proceedings, ICAASP, 1989, 242-245.

Jespersen, O. (1922) *Language, Its Nature, Origin, and Development.* (Allen and Unwin: London).

Karlsson, I. (1986) "Glottal wave forms for normal female speakers', Jour. Phonetics, 14, 415-419.

Karlsson, I. (1989) "A female voice for a text-to-speech system", Procs. Eurospeech '89, Paris, 349-352.

Karlsson, I. (1991) "Female voices in speech synthesis", Jour. Phonetics, 19, 111-120.

Karlsson, I. (1992) "Voice variations in female speech", Speech Communication, 11 (forthcoming).

Klatt, D. H. (1982) "Speech processing strategies based on auditory models", pp.181-196, in R. Carlson & B. Granström (eds.) *The Representation of Speech in the Peripheral Auditory System,* (Elsevier: Amsterdam).

Klatt, D.H. (1986) "Detailed spectral analysis of a female voice", Jour. Acoust. Soc. Am. 80, S97.

Klatt, D. H. (1987) "Acoustic correlates of breathiness: first harmonic amplitude, turbulence noise, and tracheal coupling", Jour. Acoust. Soc. Am. 82, S91.

Klatt, D.H. & Klatt, L.C. (1990) "Analysis, synthesis, and perception of voice quality variations among female and male talkers", Jour. Acoust. Soc. Am. 87, 820-855.

Lakoff, R. (1975) *Language and Woman's Place.* (Harper & Row: New York).

Markel, N.N., Prebor, L.D. & Brandt, J.F. (1972) "Biosocial factors in dyadic communication: sex and speaking intensity", Jour. Pers. and Soc. Psych., 23, 11-13.

McConnell-Ginet, S. (1983) "Intonation in a man's world", pp.69-88, in B. Thorne, C. Kramarae, & N. Henley (eds.) *Language, Gender, and Society,* (Newbury House: Rowley, Mass.).

Murry, T. & Singh, S. (1980) "Multidimensional analysis of male and female voices", Jour. Acoust. Soc. Am. 68, 1294-1300.

Pellowe, J. & Jones, V. (1978) "On intonational variability in Tyneside speech", pp.101-121, in P.Trudgill (ed.) *Sociolinguistic Patterns in British English,* (Arnold: London).

Pickering, J.B. (1988) "Glottal pulse shapes, naturalness, and the synthesis of female speech", Proceedings, 7th. FASE Symposium, Edinburgh, 1107-1114.

Tielen, M. (1992) *Male and Female Speech: An Experimental Study of Voice and Pronunciation Characteristics.* Ph.D. Dissertation, University of Amsterdam.

Titze, I.R. (1988) "Physiologic and acoustic differences between male and female voices", J. Acoust. Soc. Am. 85, 1699-1707.

Woods, N. (1992) "It's not what she says, it's the way that she says it: the influence of speaker-sex on pitch and intonational patterns", Paper presented at Intl. Workshop on Prosody, Univ. Pennsylvania, PA, 5-12 August.