A NOVEL EFFICIENT ALGORITHM FOR VOICE GENDER CONVERSION

Bob Lawlor and Anthony D. Fagan
University College Dublin, Ireland

ABSTRACT

Realistic Voice Gender Conversion (VGC) requires independent scaling of the glottal (pitch) and vocal tract (formant) related features of the input speech signal. We present a VGC algorithm which has two novel features. Firstly, an efficient frequency scaling algorithm is presented. Secondly, we use this to scale all frequencies in the input signal by the desired formant scaling factor. We then deconvolve the glottal contribution using a standard linear predictive analysis and frequency scale it further such that the desired pitch scaling factor is equal to the product of the two frequency scaling factors. Finally, we resynthesize the converted speech. The female-to-male results were excellent while the male-to-female results sounded synthetic.

1. INTRODUCTION

Voice Gender Conversion (VGC) consists of modifying female (male) speech such that it sounds male (female). Applications include voice gender normalisation for improved speech compression or recognition and voice disguise. In practice, such applications require that the VGC algorithm runs in realtime on inexpensive hardware. The human auditory system is highly sensitive to voice perception and synthetic speech, though intelligible, often sounds unnatural. The VGC challenge is to convert the gender related parameters of the speech signal without affecting naturalness. For a long time it was felt that pitch was the dominant cue in voice gender perception. However, Childers [1] showed that grouped formant information gave a higher automatic gender distinction success rate than pitch information. Hence, realistic VGC requires independent modification of the glottal (pitch) and vocal tract (formant) related features of the source speech signal. Atal [2] presented a VGC algorithm using linear predictive analysis to deconvolve the glottal and vocal tract contributions. He applied independent scaling factors to the pitch frequency and the formant frequencies and bandwidths prior to resynthesizing to give gender converted speech. The scaling factors were based on typical male-to-female vocal cord membrane and vocal tract length ratios. Our approach is similar to Atal's, but has two novel features. Firstly, we present an efficient frequency scaling algorithm based in the principle of time-domain overlap-add (TD-OLA). Secondly, we use this to scale all frequencies in the input signal by the desired formant scaling factor. We then deconvolve the glottal contribution using Atal's linear predictive analysis and frequency scale it further such that the desired pitch scaling factor is equal to the product of the two frequency scaling factors. Finally, we resynthesize the converted speech.

2. VOICE GENDER DIFFERENCES

A detailed explanation of the speech production mechanism can be found in [3]. A typical male vocal tract is about 17.5 cm in length (i.e. from vocal cords to lips) while that of a female is about 15.2 cm. The adult male larynx is about 1.2 times the size of that of the female. During puberty the male larynx undergoes a change in shape (Adam's apple protrusion) such that the adult male vocal cord membrane length reaches about 1.6 times that of the female. Analysis of male and female voiced utterances shows that the female formant frequencies are about 15% higher than male. This difference is in close agreement with the male-to-female vocal tract length ratio. Female pitch is generally about 1.7 times that of male. This difference is attributed mainly to the difference in vocal cord membrane length although other factors such as male/female differences in the way in which the cords open and close are believed to be relevant also [4]. The air pressure variation produced in the region of the vocal cords is known as the Glottal Volume Velocity Waveform (GVVW). During voicing this has the form of that shown in Figure 1 [5].

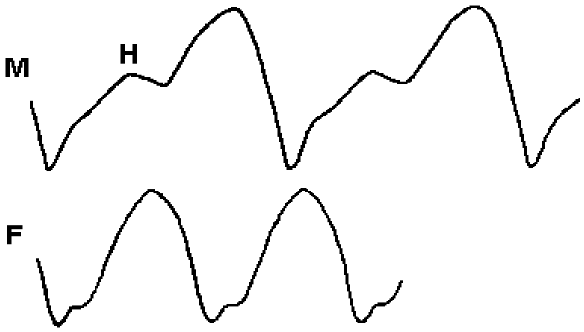

Figure 1. Typical male (**M**) and female (**F**) GVVWs.

Monsen [5] suggests, and cites supporting references, that the hump (indicated **H**) in the opening phase of the male GVVW may be due to a slightly out of phase movement of the upper and lower parts of each vocal cord, whereas the shorter female cords come into contact with each other more as a single mass. Four parameters which are often used to characterise the GVVW are:

1. The AC flow which is equal to the peak-to-peak value of the GVVW as shown in Figure 1.
2. The Maximum Flow Declination Rate (MFDR) is defined as the minimum of d/dt(GVVW) and is generally equal to the slope of the falling edge of the glottal pulses.

3. The pitch period, T, which is the time between sucessive peaks of the GVVW. The pitch frequency, F_0 is equal to 1/T.
4. The average flow (Avg. Flow) which is equal to the mean of the GVVW.

Sulter [6] carried out extensive tests on GVVW characteristics across gender for both vocally trained and untrained subjects under soft, normal and loud voicing conditions. Table 1 summarises the parameters of the GVVW which showed major differences across gender for the untrained subjects under normal voicing conditions only.

Parameter	Female	Male	Ratio (F/M)
AC flow (*l/s*)	0.26	0.57	0.46
MFDR (l/s^2)	504	1026	0.49
1/T (Hz)	253	145	1.7
Avg flow (*l/s*)	0.16	0.23	0.7

Table 1. GVVW parameters most affected by gender

2.1. Voice gender perception

For a long time it was believed that pitch was the dominant gender cue. However, recent studies of the speech production mechanism have revealed more subtle differences between features of the male and female speech waveforms. One such study [1] showed that grouped formant information can provide a slightly higher automatic gender distinction success rate (98.1%) than pitch information (96.2%). These figures suggest that both pitch and formant information are important cues in voice gender distinction.

3. LINEAR PREDICTIVE ANALYSIS/SYNTHESIS

Dudley [7] presented the *channel vocoder* (VOice CODER) for speech compression. This is based on modelling the vocal tract response with a time-varying filter. He used a 10 band filterbank at the encoder to sample the short time magnitude spectrum. At the decoder, these magnitude estimates were used to control a bank of oscillators corresponding to the filterbank centre frequencies. Flanagan [8] presented a modification to the channel vocoder called the *phase vocoder*. This included samples of the phase spectrum (as well as samples of the speech signal short time magnitude spectrum) which contains time-varying pitch (and its harmonics) information. These phase samples were used at the decoder to control the phase of the oscillators. Flanagan found that using the phase vocoder in a back-to-back analysis/synthesis system gave output '*quality much better than the conventional channel vocoder*'.

Atal [2] presented the linear predictive approach to speech coding based on the same speech production model as that used by Dudley and Flanagan, i.e. a source excitation signal driving a time-varying filter. However, Atal applied a time domain least-mean square analysis algorithm directly to the speech signal to estimate the parameters of the time-varying vocal tract model filter. By passing the input speech signal through the inverse of the vocal tract model filter he produced a residual signal which was a good approximation to the glottal source excitation signal. By passing this residual signal back through the same vocal tract model filter, the original speech is perfectly reconstructed. This analysis/synthesis approach forms the basis of many modern speech compression algorithms e.g. Code Excited Linear Prediction (CELP), Multi-Pulse Excited Linear Predictive Coding (MP-LPC) [9]. The spectral peaks corresponding to the formant frequencies are perceptually most important. As such peaks can be accurately modelled using poles, Atal chose an all-pole vocal tract model filter. The all-pole model also simplified the filter parameter estimation. Atal's main application of interest was efficient storage and transmission of speech. However, he also details a number of other applications of his linear predictive analysis/synthesis algorithm such as formant analysis and separating the spectral envelope and fine structure. He notes that '*the synthesis procedure allows independent control of such speech characteristics as spectral envelope, relative duration's, pitch and intensity*'. One such application which he presents under the heading of '*re-forming the speech signal*', involves simulating a female voice from parameters derived from a male voice. For this application he used a 10-th order model filter (predictor). He scaled the pitch period by a factor of 0.58 and the formant frequencies by 1.14. He also scaled the formant bandwidths by a frequency-dependent factor given by 2 – F_i/5000 where F_i is the formant centre frequency in Hz. The 5000 in this factor is equal to half his speech signal sampling frequency.

4. TIME & FREQUENCY SCALING

If a signal is sampled using sample frequency, f_s samples per second, and then played back using a different conversion frequency, f_p, the duration of the signal will be scaled by the factor, f_s/f_p and the frequency content of the signal will be scaled by the factor f_p/f_s. Time-Scale Modification (TSM) of a speech (or other audio) signal consists of modifying its duration without affecting its perceived frequency content. Similarly, Frequency-Scale Modification (FSM) consists of modifying its frequency content without affecting its duration. In 1985 Roucos [10] presented the Synchronised Overlap-and-Add (SOLA) algorithm for speech TSM. With this approach, overlapping segments (or frames) of the input signal are first extracted, a frame being typically several pitch periods in duration. By decreasing the overlap between successive frames, time-scale expansion is realised. Similarly, by increasing the overlap, time-scale compression is realised. In the original SOLA algorithm [10], the segment alignment was optimised by computing a normalised cross-correlation measure, *R*, for a range of possible alignment offsets and then choosing the offset for which *R* is a maximum, indicating maximal similarity between overlapping segments. High quality combined with moderate computational load has made the SOLA algorithm the choice for many speech and audio TSM systems. We present an alternative TSM algorithm which offers a significant reduction in computational load without loss of quality. If a signal is time-scaled by some factor, T_S, and then played at T_S times its original sample rate, the net effect is to scale the frequency content by the factor, T_S, without affecting the original duration, i.e. FSM. We use this approach to frequency scale both the glottal and vocal tract related components of speech independently to realise voice gender conversion. First we describe our efficient TSM algorithm which we call Adaptive Overlap-Add (AOLA).

4.1 Adaptive Overlap-Add (AOLA)

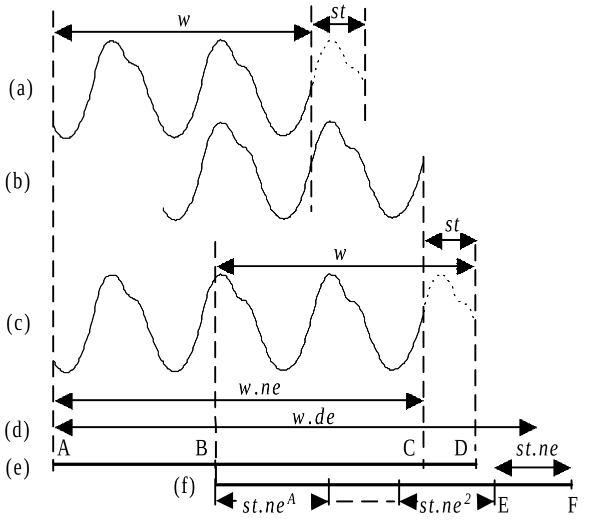

Figure 2. Adaptive Overlap-Add Algorithm

Referring to Figure. 2, the solid trace of plot (a) represents a rectangular windowed segment of the input signal. The window length, *w*, is chosen such that it will accommodate at least two pitch periods of typical male voiced speech. For unvoiced speech, the choice is not critical and *w* can be left equal to the value chosen to satisfy the above voiced condition. Assuming we wish to scale the duration of this segment by some desired expansion factor, *de*, the steps involved in the algorithm are as follows:

1. The windowed input segment (a) is duplicated and the duplicate aligned with (a) as shown in plot (b). The alignment criterion is based on aligning the two largest peaks or troughs.
2. A synthetic segment, (c), is produced by fading gradually from (a) to (b) in the overlapping region. The natural expansion factor, *ne*, is given by the ratio of the lengths of (c) and (a) as indicated.
3. The rectangular window is stepped forward in time by *st* = |CD| = *w.(1-ne)/(1-de)* and the new step-size segment of the original concatenated with (c) as indicated dotted such that the next segment to be expanded is the length *w* portion of (c) above BD, see plot (e). Repeat from step 1 until the end of the signal being scaled is reached.

Rationale: Plot (d) represents the desired length to which we wish to time-scale (a), i.e. *w.de*. The segment of (c) above AB has been time-scaled by the desired factor, *de*, and is output from the time-scaling window. For each step of the window we repeat the peak search and update *ne*. Assuming *ne* to be approximately equal to its last value, segment BD of (c) expands in the same way as (a) to length |BF| ≈ *w.ne*. Step size portion CD of (c) expands to length |EF| ≈ |CD|.*ne*, but we require it to be expanded by factor *de*. To achieve this we must apply our natural expansion factor *A* times such that ne^A = *de*. If segment CD is to have *A* applications of natural expansion factor, *ne*, before leaving the expansion window, then from plot (f), our step size, *st*, must satisfy the following equation

$$st.ne + st.ne^2 + \cdots + st.ne^A \approx w.ne \qquad (1)$$

$$\Rightarrow \quad st \approx w.\frac{1 - ne}{1 - ne^A} = w.\frac{1 - ne}{1 - de} \qquad (2)$$

As *ne* is continuously varying, (1) (and (2)) is an approximation. In fact, each of the *ne* and *st* terms in (1) are slightly different. By updating *ne* and hence *st* for every step of the window, the algorithm accurately adapts to the local signal characteristics.

For time-scale compression the approach is similar. In this case the peaks or troughs are aligned as before but the sections of (c) to the left and right of the central overlapping section are discarded leaving a naturally compressed segment. If the input segment has a natural compression factor, *nc*, and the desired compression factor is *dc*, then equation (2) becomes

$$st = w.\frac{1 - nc}{1 - dc} \qquad (3)$$

5. VOICE GENDER CONVERSION

For female-to-male VCG we require to scale the formant frequencies by some factor, S_F, (~0.87 i.e. down 15%) and the pitch frequency by some factor, S_P, (~0.6). The steps involved are as follows:

1. Time-scale modify the input signal by the factor S_F using the AOLA algorithm. This shortens its duration without affecting its frequency content. If it is played back now at S_F times its original sample rate, the duration will be restored to its original but all frequencies (pitch and formants) lowered by the factor S_F.
2. Apply linear predictive inverse filtering to separate the formant information form the spectral fine structure (pitch).
3. Time-scale modify the inverse filter output signal from step 2 (residual) by the factor S_P/S_F using the AOLA algorithm.
4. Using the vocal tract model filter coefficients of step 2 and the time-scale modified residual of step 3, resynthesize the speech signal and play it back at S_P times its original sample rate.

For male-to-female VGC the procedure is identical, but the scaling factors are equal to the reciprocals of those for female-to-male.

6. RESULTS

We applied the algorithm to a selection of signals from the DARPA TIMIT speech corpus [11]. We present here sample results. The full path names of these speech signals within the corpus are:

\TIMIT\TEST\DR1\FELC0\SA1.WAV: for the original female speech sample [SOUND 0027_01.WAV] and

\TIMIT\TEST\DR1\MDAB0\SA1.WAV: for the original male [SOUND 0027_02.WAV].

We tried various combinations of pitch and formant scaling factors and filter orders. For female-to-male the pitch scaling factors, S_P were in the range 0.55 to .8 and the formant scaling factors, S_F in the range 0.82 to .92. The filter orders tried were 6, 8, 10 and 12. A filter order of 8 gave better results than 6, while 10 and 12 gave no improvement over 8 so we used 8. We found that $S_P = 0.74$ and $S_F = 0.87$ gave the best female-to-male result [SOUND 0027_03.WAV]. For male-to-female [SOUND 0027_04.WAV] we used the same filter order and the reciprocal scaling factors. We also generated back-to-back results, i.e female-to-male-to-female [SOUND 0027_05.WAV] and male-to-female-to-male [SOUND 0027_06.WAV]. These results are also available from the author by email (rlawlor@faraday.ucd.ie).

7. DISCUSSION

The female-to-male results were excellent, the output sounding completely natural. The male-to-female results were not as good with many of the output samples sounding unnatural. These findings are consistent with those of other VGC researchers [12] [13]. Our scaling factors were very close to those used in [12]. Childers [12] suggests that the reason why male-to-female VGC is less successful when a single formant scaling factor is used is that the factor should itself be frequency dependent (‘*higher frequency formants should be shifted less than lower frequency formants*’). As the single compromise factor is greater than unity for male-to-female and less than unity for female-to-male, the inadequacies of the compromise will be more noticeable on male-to-female. Chan [13] cites [12] but suggests that the reason male-to-female VCG is more difficult is ‘*unknown at present*’. For the improved speech recognition application and for most voice gender disguise applications female-to-male VGC alone is adequate. The improved speech compression application requires good back-to-back performance. The back-to-back outputs sounded like slightly noisy versions of their originals but suggest that the algorithm can be used to normalise voice gender to give greater compression of telephone quality speech.

8. CONCLUSION & FURTHER WORK

We have presented an efficient VGC algorithm which gives high quality female-to-male VGC but poor quality male-to-female VGC. We feel that further improvement may be possible with a suitable non-linear transformation of the GVVW. This would, however, increase the computational burden. The algorithm presented can be implemented in real-time on a low cost digital signal processor.

REFERENCES

[1] Childers, D. G. and Wu, K. 1991. Gender recognition from speech. Part II: Fine analysis. *Journal of the Acoustical Society of America,* Vol. 90, pp. 1841 – 1856.

[2] Atal, B. S. and Hanauer, S. L. 1971. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America,* Vol. 50 No. 2, pp. 637 – 655.

[3] O’Shaughnessy, D. 1987. *Speech Communication : Human and Machine.* Addison – Wesley.

[4] Titze, I. R. 1989. Physiologic and acoustic differences between male and female voices. *Journal of the Acoustical Society of America,* Vol. 85, No. 4, pp. 1699 - 1707.

[5] Monsen R. B. and Engebretson, A. M. 1977. Study of variations in the male and female glottal wave. *Journal of the Acoustical Society of America,* Vol. 62, No. 4, pp. 981 - 993.

[6] Sulter, A. M. and Wit, H. P. 1996. Glottal volume velocity waveform characteristics in subjects with and without training, related to gender, sound intensity, fundamental frequency and age. *Journal of the Acoustical Society of America,* Vol. 100, No. 5, pp. 3360 - 3373.

[7] Dudley, H. 1939. Remaking speech. *Journal of the Acoustic Society of America,* Vol. 11, No.2, pp. 169 - 175.

[8] Flanagan, J.L. and Golden, R.M. 1966. Phase Vocoder. *Bell system technical Journal.*

[9] Deller, J. R. JR., Proakis, J. G. and Hansen, J. H. L. 1993. *Discrete-Time Processing of Speech Signals.* Macmillan Publishing Company.

[10] Roucos, S. and Wilgus, A. M. 1985. High-quality time-scale modification for speech. *IEEE proceedings on acoustics, speech and signal processing.*

[11] DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, http://www.ntis.gov/fcpc/cpn4129.htm

[12] Childers, D. G., Wu, K., Hicks, D.M. and Yegnanarayana, B. 1989. Voice conversion. *Speech Communication* 8, pp. 147-158.

[13] Chan, P. A. and Damper, R. I. 1994. Voice conversion by whole-spectrum scaling. *ESCA Workshop on automatic speaker recognition, identification and verification,* pp. 165 - 168.