Implementation of Real–Time AMDF Pitch–Detection for Voice Gender Normalisation

E. Jung
Dublin Institute of
Technology, Kevin St.,
Dublin 8, Ireland
elmar.jung@dit.ie

A. Schwarzbacher
Dublin Institute of
Technology, Kevin St.,
Dublin 8, Ireland
andreas.schwarzbacher@dit.ie

R. Lawlor
National University
of Ireland
Maynooth, Ireland
rlawlor@eeng.may.ie

4th March 2002

Abstract

Traditionally the interest in voice gender conversion was of a more theoretical nature rather than founded in real-life applications. However, with the increase in mobile communication and the resulting limitation in transmission bandwidth new approaches to minimising data rates have to be developed. Here voice gender normalisation (VGN) presents a novel method of achieving higher compression rates by using the VGN algorithm to remove all gender specific components of a speech signal and thus leaving only the information content to be transmitted.

A second application for VGN is in the field of speech controlled systems, where current speech recognition algorithms have to deal with the voice characteristics of a speaker as well as the information content. Here again the use of VGN can remove the speakers voice characteristics leaving only the pure information. Therefore, such a system would be capable of achieving much higher recognition rates while being independent of the speaker. This paper presents the theory of a gender removal system based on VGN and furthermore, outlines an efficient real-time hardware implementation for the use in portable communications equipment.

1 Introduction

Voice normalisation is a main focus of the research on voice conversion systems. Voice compression and speech recognition perform significantly better if the range of speakers is limited to known subjects. The reason is an effective limitation of the signal space that has to be analysed, which allows the system to analyse less parameters.

A successful implementation of a gender–dependent speech compression system is described by Marston [1]. After a gender detection step the speech is compressed using a dedicated speech encoder for the found voice gender. The gender detector is using solely the pitch information of the speech wave and is cepstral based. The outcome of this method is a robust system with a bit–rate reduction at maintained speech compression quality.

Vergin [2] presents a robust continuous speech recognition system is that uses gender–dependent modelling. This implementation also firstly applies a voice gender detection before the actual gender specific processing. The performance of the gender–dependent system shows a 14 % improvement in the correct recognition rate compared to the speaker independent recognition system.

Several suggested voice conversion methods are based on the idea of vector quantisation and mapping, like proposed by Childers [3]. In these systems the speech wave is analysed into subspaces and a codebook with analysis vectors is generated. The actual transformation is then the problem of converting analysis vectors of the source speaker to the corresponding vectors of the target speaker prior to synthesising the modified speech. This approach works quite well for source–target conversion, even across voice gender. However, the size of the codebook is an inherent limitation in possible conversions and conversion quality. It is therefore desirable to achieve a general, non–personalised voice transformation with the focus to voice normalisation.

Lawlor [4] introduces a voice gender conversion (VGC) method based on a parametric source–filter model that is suitable for voice gender normalisation. The gender normalisation is done by converting female voice into the male signal spectrum. Since the voice transformation is reversible this system is well suitable for voice normalisation. This paper presents an refined version of the mentioned VGC algorithm, as well as the hardware implementation of an AMDF pitch detector with a voiced/unvoiced decision instance.

2 Methodology

The proposed VGC performs the voice conversion adapting the conventional source–filter model of speech as presented by Fant [5]. The source–filter description is a very accurate model of the human speech production mechanism as it is closely related to the physiological features of the vocal tract. Gender–specific features are strongly related to these

0–7803–7503–3/02/$17.00 ©2002 IEEE

physiological properties. The human speech production mechanism consists of the larynx and the vocal tract. For voiced utterances the larynx produces the periodic excitation or source signal, while the vocal tract acts as a time–varying resonator or filter for this signal. The source signal is characterised by the fundamental or pitch frequency F_0, while the vocal tract filter is described by the centre frequencies of its resonant peaks, the formants. In the frequency domain representation the spectrum of a voiced utterance is in principle a line spectrum with harmonics of the pitch frequency, which is shaped by the spectral envelope of the formants.

The first four formants F_1–F_4 are characteristic for the description of voiced utterances. Unvoiced utterances originate from noisy or aperiodic excitation, such as turbulent air flow. Since their energy content is small compared to voiced utterances they do not carry significant speaker dependent information and are therefore not considered for voice conversion.

The pitch signal is generated in the larynx during voiced speech and its frequency depends mainly on the physical size of the larynx. Since the male larynx is in average 1.7 times larger than the female equivalent its pitch frequency is 1.7 times lower. The vocal tract length and shape is responsible for the formant structure. The male vocal tract is in average 1.2 times longer resulting in a formant spectrum which is 1.2 times lower. These figures are average values from a fundamental study by Childers [6]. However, a fine analysis shows, that the pitch frequencies of male and female voices have a linear relationship at a ratio of 1.76, while the formants exhibit a non–linear correspondence. The first three formants F_1–F_3 have a linear relationship across the two voice genders with an average ratio of 1.2, while the fourth formants F_4 has a slightly smaller ratio of 1.18, as seen in Table 1.

Conversion	F_0	F_1	F_2	F_3	F_4
female/male	0.57	0.83	0.83	0.83	0.85
male/female	1.76	1.20	1.20	1.20	1.18

Table 1: Average Pitch and Formant Scaling Factors for Female to Male Voice Gender Conversion, calculated from Data evaluated by Childers [6]

The principle of the presented VGC is to separate source and filter spectrum and independently frequency–scale the pitch and formant frequencies to achieve a spectral transformation between the female and the male voice spectra. Female voice spectra exhibit a broader range of variance and are therefore more difficult to model. Therefore the male voice gender spectrum was chosen for the gender normalisation of speech. Since speaker dependent and hence gender dependent information is carried mainly in voiced utterances the VGC system focuses on the transformation of voiced speech segments. This is also with respect to computational efficiency.

The relationships between male and female pitch and formant frequency spectra have been evaluated by Childers [6] for a range of representative voiced utterances. The findings of this study show a linear relationship for the pitch frequency with a scaling factor of 1.76, and a slightly non–linear relationship for the first four formant frequencies. The first three formants have an average scaling factor of 1.2, while the fourth formant requires a smaller scaling of only 1.18. These scaling factors hold for the conversion of a male voice into a female.

3 Voice Gender Conversion using AOLA

To follow the outlined idea of VGC the pitch and formant spectra have to be shifted by different scaling factors. The frequency shift is achieved using a TSM algorithm called adaptive overlap and add (AOLA), as presented by Lawlor [7]. Overlap and add TSM methods allow to time–scale a signal without affecting the frequency contents of the signal. However, if an expanded signal is resampled or played back at a lower playback rate to match the original duration of the signal the frequency contents is shifted linearly by the expansion factor. AOLA is an efficient new algorithm that reaches the quality of the commonly used synchronised overlap and add (SOLA), but has a computational burden an order of magnitude less than SOLA.

In case of male to female conversion this is done by applying TSM to expand the entire signal by the formant scaling factor S_F. If the signal were now resampled to its original length the frequency spectrum and with it the formants would have been shifted by the TSM expansion factor. Then the signal frames are LP analysed in case they are voiced or unvoiced. Voiced frames are LP analysed and inversely filtered to separate pitch and formants. The pitch information is inherent in the LP residual. The residual is TSM scaled by $\frac{S_P}{S_F}$, where S_P is the required pitch scaling factor, to compensate for the overall scaling during the formant scaling. The time–scaled residual is then used with the LPA filter to resynthesise the modified signal frame. However, unvoiced frames do not contain much speaker information and therefore do not need to be LP analysed. These frames are directly scaled by $\frac{S_P}{S_F}$, thus saving the LPA and filtering steps. The overall converted signal is generated by concatenating the modified frames.

The necessary frequency shifts of the pitch and formant spectra are achieved using the TSM algorithm AOLA. The two major signal components, pitch and formant spectrum, require both different scaling factors, as pointed out above. Furthermore, the formant spectrum exhibits a non–linear behaviour as F_4 needs to be scaled less than the three lower formants. This is taken into account by refining the model to accommodate this non–linearity.

This problem is approached by linearising the relationship between male and female formant spectra. This is basically done by splitting the speech signal into two complementary frequency bands to separate F_4 from the other formants. The low–pass signal contains the pitch and its harmonics and F_1–F_3, while the high–pass signal contains higher harmonics of the pitch and F_4.

The two signals are then processed in the same way as described above, with the difference that different formant scaling factors are applied to each subband. The low-pass band is scaled by $S_{F\,low} = 1.2$, while the high-pass band requires scaling by $S_{F\,high} = 1.18$. After the individual processing the modified signal is composed by merging the two individually scaled subbands.

The cut–off frequency for the band splitting filter is 3 kHz for male to female conversion (3.5 kHz respectively for the other direction). With 350 Hz the transition band is fairly wide, which allows the use of a low filter order. The implementation uses a fourth order elliptic filter with a linear phase response, as described by Oppenheim [8]. The linearity of the filter is essential in order to avoid phase distortion, which would degrade the signal quality.

3.1 AMDF Pitch Detection

Pitch detectors using average magnitude difference function (AMDF) are computationally very efficient. Other reliable methods of pitch detection use the autocorrelation or cross–correlation function. However, the latter methods require multiply functions which take up a large area in a hardware implementation. The structure of the implemented AMDF pitch detector is shown in Figure 1.

3.2 Application of VGC to Voice Gender Normalisation

In order to apply voice gender normalisation to a speech signal firstly the voice gender has to be detected. This done based on the pitch information. Female and male voices show a significantly different pitch contour [6], which was found to be sufficient to make an accurate voice gender decision after a short time of speech analysis. In case a female voice is detected the VGC algorithms will be invoked and the voice will be transformed into the male voice spectrum.

The processing of the input signal is as described above with the difference that the pitch AMDF detection is applied before the VGC section. Based on the detected pitch the voice gender is estimated. In case a female voice is sensed the voice will be normalised using the VGC system. If a male voice is detected the VGC will be bypassed, as the desired spectral properties are already present. The normalised voice can then be further processed in a speech compression or recognition system. In the first case the gender normalised voice has to be denormalised after transmission and decompression. The denormalisation is achieved by applying the same VGC steps as for the normalisation, but using the reciprocals of the scaling factors.

4 Experiments and Results

The computational performance of this refined VGC algorithm has been successfully tested for a range of common sampling rates ranging from 8 kHz to 20 kHz. The LPA filter order was modified between 6, 8, 10 and 12. While an 8th order filter showed improved performance over a 6th order filter, a further increase in the filter order did not result in a further improvement.

Table 2 shows the results of a listening test that was carried with the improved VGC in a comparison with the original VGC. The speech data originate from the TIMIT DARPA [9] speech corpus of American English and were sampled at 20 kHz. For the test the signals were downsampled to 8 kHz to use a sampling rate that is in the range of possible target applications such as telephony. The results indicate that the performance has effectively improved using the refined model.

The Speech Sample sounds ...		
Better	Equal	Worse
	than Reference	
Female to Male Conversion		
44%	50%	6%
Male to Female Conversion		
19%	75%	6%
Female to Male to Female Conversion		
25%	75%	0%
Male to Female to Male Conversion		
19%	75%	6%

Age/Gender	Male	Female
0–10		
11–20	5	2
21–30	4	3
31–40	–	1
41–50		
50+	1	–

Table 2: Results of the VGC Listening Test of the Improved VGC System and the Age/Gender Distribution of the Test Listeners

Test listeners judged the converted gender correctly, but mostly as sounding unnatural. The performance was found to be better for the female to male conversion. Figure 1 shows the block diagram of a hardware implementation of the algorithm. The main focus on the implementation of the VGN is in the field of portable equipment [10]. Therefore, the the circuit was developed to keep area and power to a minimum.

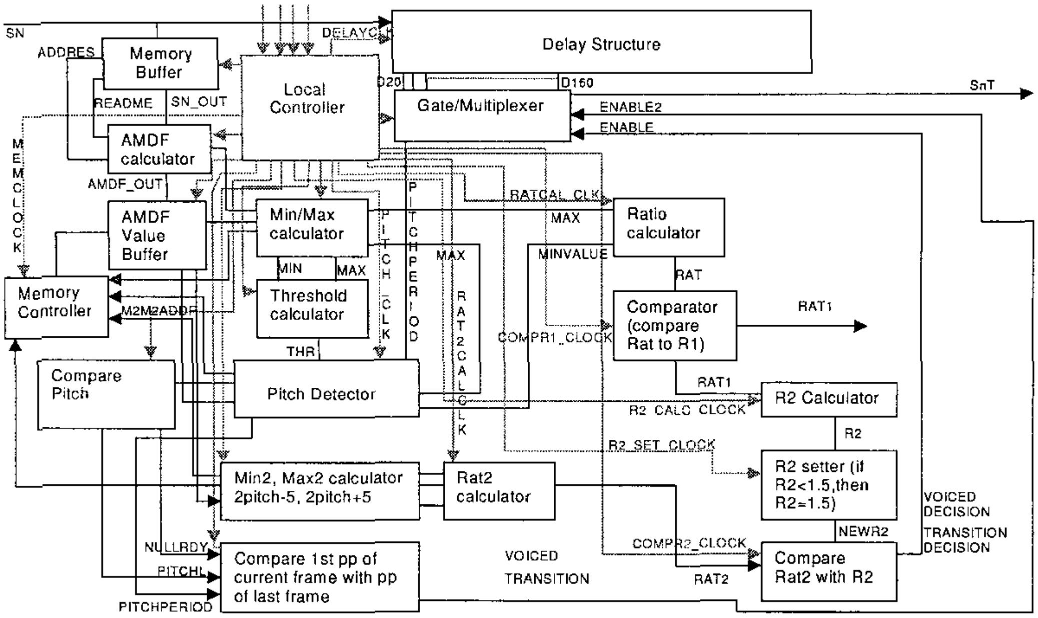

Figure 1: Schematic Diagram of the AMDF Pitch Detection Algorithm

5 Conclusions

This paper has presented a refined voice gender conversion method using the efficient time–scale modification algorithm AOLA. The VGC is suitable to achieve voice gender normalisation. For application of voice normalisation in speech compression systems it is essential that this normalisation can be reversed to restore the original signal after processing or compression. This task was successfully carried out by the proposed algorithm. However, the conversion of female to male voices is more successful than vice versa. This is due to larger variances among female voices and a resulting broader signal spectrum.

Finally, the block diagram of a hardware implementation was presented which will be used in the next stage of the research project to develop an energy efficient circuit for the use in portable communications equipment.

References

[1] D.F. Marston, 'Gender Adapted Speech Coding,' *IEEE Trans. Acoustics, Speech and Signal Processing,* Vol. 1, pp. 357–360, 1995.

[2] R. Vergin, A. Farhat and D. O'Shaughnessy, 'Robust gender–dependent acoustic-phonetic modelling in continuous speech recognition based on a new automatic male/female classification,' *Proc. International Conference on Spoken Language,* Vol. 2, pp. 1081–1084, 1996.

[3] D.G. Childers, K. Wu, D.M. Hicks and B. Yegnanarayana, 'Voice Conversion,' *Speech Communication,* Vol. 8, No. 2, pp. 147–158, 1989.

[4] R. Lawlor and A.D. Fagan, 'A Novel Efficient Algorithm for Voice Gender Conversion,' *XIVth International Congress of Phonetic Sciences,* University of California, Berkeley, USA, 1999.

[5] G. Fant, *Acoustic Theory of Speech Production.* Mouton & Co., S'Gravenhage, 1960.

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

[7] R. Lawlor and A.D. Fagan, 'A Novel Efficient Algorithm for Audio Time–Scale Modification,' *Irish Signals and Systems Conference,* National University of Ireland, Galway, 1999.

[8] A.V. Oppenheim and R.W. Schafer, *Discrete–Time Signal Processing.*
Prentice-Hall, Englewood Cliffs, New Jersey, 1989.

[9] DARPA TIMIT, *Acoustic–Phonetic Continuous Speech Corpus.*
American National Institute of Standards and Technology, NTIS order number PB91-50565.

[10] A.Th. Schwarzbacher and J.B. Foley, 'Challenges of integrated circuit design in the 21st century,' *Engineering Design in an Academic Environment, Royal Irish Academy, Dublin, Ireland,* pp. 31–32, October 2000.