A new Multi-Speaker Formant Synthesizer that applies Voice Conversion Techniques

J. M. Gutiérrez-Arriola, J.M. Montero, J.A. Vallejo, R. Córdoba, R. San-Segundo, J.M. Pardo*

Grupo de Tecnología del Habla, Dpto. de Ingeniería Electrónica, ETSIT, Universidad Politécnica de Madrid.

*Also in Dpto. de Ingeniería de Circuitos y Sistemas, EUITT, Universidad Politécnica de Madrid.

Contact: juana@die.upm.es, http://www-gth.die.upm.es

Abstract

We present a multi-speaker formant synthesizer based on parameter concatenation. The user can choose among three speakers, two males and one female. The synthesizer stores all the parameters for the basic speaker and linear transformation functions to synthesized the other two. The complete database for one speaker consists of 455 parameterized units (diphones, triphones,...) and the parameters used are pitch, formants and bandwidths and source parameters (four parameters for the LF model, and glottal noise). To get the converted speaker we store a linear transformation function for each spectral stable segment of each unit. Preliminary results show that the quality of the synthesizer is very good and that this system can help us to study and understand the speaker variability problem.

1. Introduction

One of the lacks of current synthesizers is that they are commonly judged as monotonous and boring. Recent researches try to improve prosodic models and to add variability to the voices. This variability includes emotions [1], [2], and speaker conversion [3], [4].

Speaker variability is important when the synthesizer is used in a talker machine or as a reader because users can choose the voice they like. In information retrieval systems, it has been proved that a change of the speaker is more effective when transmitting the message, if the information to provide is rather long.

We propose a synthesizer based on parameterized unit concatenation. The basic database consists of 455 units that can be phones, diphones or triphones. We extract unit parameters semi-automatically. Extracted parameters are: pitch, first five formants and bandwidths, glottal noise and four parameters of the LF model (open quotient, skewness, speed quotient and tilt) [5]. Pitch and formants are manually revised. Parameterized units are then concatenated to produce speech, smoothing formants at the transitions.

Voice transformation is performed in a unit-by-unit basis and only for voiced speech. Each unit from the basic speaker is aligned with the same unit from the desired speaker. For a given parameter a correspondence between the original and the desired speaker is obtained and it is approximated by a linear function using a linear regression algorithm [6].

When a different voice is selected in the synthesizer the coefficients of the linear function are applied to each parameter to get the desired voice. Results confirm that the resulting voice is more similar to the desired speaker than to the basic one.

2. Analysis

The 455 units for the three speakers are first pitch marked with an algorithm very similar to that defined in [7]. Then, they are pitch synchronously analyzed using Durbin algorithm to calculate the linear prediction coefficients (LPC). The analysis window that we used is a two-period long Hamming window centered on every pitch mark. The original signal is filtered using these coefficients to obtain the LP excitation signal or LP residual.

The coefficients LP will be used to calculate formants and bandwidths, while the LP excitation signal will represent the second derivative of the glottal flow, and its integral will be approximated by an LF model to get the source parameters.

2.1. Formants and Bandwidths

The first five formants and bandwidths have to be estimated from the LP polynomial. One method for estimating formants is to factor the LP polynomial and to assign the appropriate roots to simulate the resonances of the vocal tract. For our analysis, a twenty-first order LP polynomial provides twenty roots, these roots must be real or complex conjugated pairs. A formant estimation procedure is applied to find which of these roots belong to the vocal tract. For each root z with angle ϕ and radius r in the z-domain, its transfer function is given by:

$$H(z) = \frac{1}{1 - re^{jf} z^{-1}}$$

If the sampling frequency is Fs (16kHz in our case), the corresponding frequency and bandwidth are defined as follows:

$$F = \frac{f}{2p} Fs$$

$$Bw = \frac{Fs}{p} \cos^{-1}\left(\frac{4r - 1 - r^2}{2r}\right)$$

After calculating frequencies and bandwidths for all the roots a formant selection algorithm is applied:

- Real roots are not taken into consideration.
- Roots that give bandwidths over 1280Hz are eliminated.
- Roots with a F/Bw relation over 0.8 are also eliminated.
- If two roots are separated less than 300Hz, the one with the greatest bandwidth is eliminated.
- If, at this step, there are more than five formants, the roots with the greatest F/Bw relation are eliminated.

To avoid artifacts and bad formant trajectories, after the selection step another algorithm is applied to smooth and correct formant trajectories.

After extracting formants and bandwidths, all the units are resynthesized and listened to; if there is any problem with the results, a manual revision of formants and bandwidths is performed for that unit.

2.2. Source parameters

2.2.1. LF parameters

The source model used by the synthesizer is the LF model as described in figure 1.

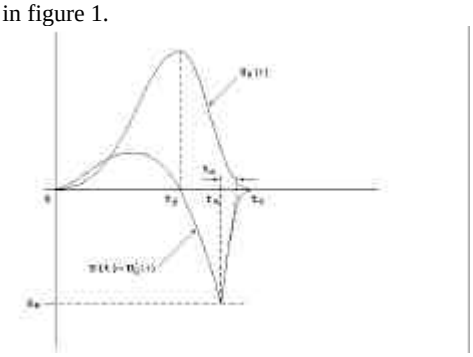

Figure 1: LF model

Where Ug(t) is the glotal flow and E(t) is the glottal flow derivative. The model is divided into two parts:

$$\frac{dU_g(t)}{dt} = E(t) = E_0 \cdot e^{at} \cdot \text{sen}(w_g t) \qquad 0 < t \le t_e$$

$$E(t) = -\frac{E_e}{et_a} \cdot [e^{-e(t-t_e)} - e^{-e(t_c-t_e)}] \qquad te < t \le t_c \le T_0$$

We need four parameters to characterize a glottal flow period [8]. These parameters are:

- Open quotient (OQ) is the portion of the pitch period during which the glottis remains opened.

$$OQ = \frac{t_c}{T_0}$$

- Speed quotient, defined as the relation between the opening phase and the closing phase.

$$SQ = \frac{tp}{tc - tp}$$

- Glottal pulse skewness.

$$Skew = \frac{t_e - t_p}{t_p}$$

- The effect of the return phase, t_a, on the source spectrum is approximately a first order low-pass filter with a cutoff frequency:

$$F_a = \frac{1}{2pt_a}$$

The parameter F_a is directly related to the spectral tilt of the glottal source.

With these four parameters we can extract all the timing parameters for the LF model. The rest of parameters can be deduced from timing parameters or they are amplitude constants that can be fixed to a desired value.

To extract these parameters we performed a two step analysis. First, we use a 6th-order polynomial waveform model to represent the derivative of the glottal volume velocity waveform [9],[10]. This derivative function is computed by direct integration of the residual and high pass filtering, in order to zero-center the signal.

The polynomial function is obtained by curve fitting in a least square sense, where a fine-tuning or readjustment is needed to exactly synchronize the pitch marks with the most negative sample.

After the polynomial fitting we adjust the LF model to the polynomial function searching the curve to find timing references and adjusting the rest of parameters to fit the polynomial in a least square error sense.

2.2.2. Glottal noise

After the LF model is constructed, the glottal noise is added to the source signal.

This noise is calculated as a gain that is proportional to the error between the integral of the LP residual and the LF approximation.

3. Synthesizer

The synthesizer received as input a list of units with their required pitch and duration (as shown in table 1 for the utterance *"la casa")*.

Unit	Duration (ms)	Final Pitch (Hz)
_LA	135	81
AK	45	81
KA	150	108
AS	70	100
S	120	100
SA	85	100
A_	95	103

Table 1: Example of the input file to the synthesizer

This file is generated with a text-to-unit application developed in the Speech Technology Group with the prosody model adapted to one of the male speakers.

The synthesizer loads the parameterized units, changes its pitch and duration according to the requirements and concatenates them to produce the desired speech.

Prosody modification is accomplish in two steps. Each parameter frame for a unit corresponds to a pitch period. The first parameter of the frame is the fundamental frequency that can be directly changed. A straight line is constructed between initial and final pitch, and the corresponding fundamental frequency is assigned to each frame. Duration is adjusted repeating or deleting frames.

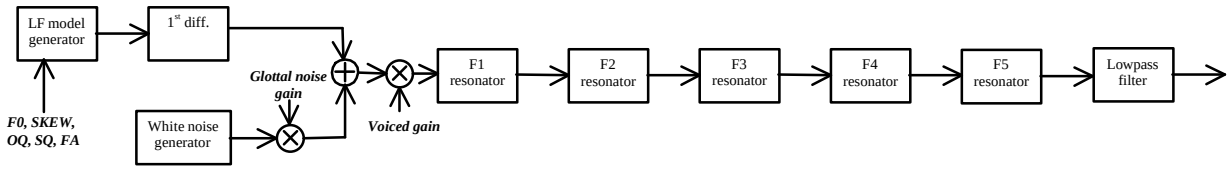

Figure 1: Formant synthesizer structure

A formant-smoothing algorithm is applied at unit transitions to avoid artifacts due to bad formant trajectories.

A formant synthesizer converts the resulting frames into speech. Figure 2 shows the synthesizer scheme for voiced sounds; a parallel structure is also present in order to synthesize unvoiced sounds, but it has been omitted.

4. Voice Conversion Algorithm

The voice conversion algorithm is described in [6]. The inputs to the algorithm are the utterances of the source and target speakers. In our case the source speaker is the basic one and the target speaker is the new one.

A dynamic time warping (DTW) algorithm is used to find the correspondence between the target-speaker and source-speaker timing. After DTW, each parameter is converted independently by means of a linear transformation of the form X=AY+B, where X are the target parameters and Y are the source parameters. A and B are calculated by means of a linear regression algorithm. The conversion parameters are: gain contour, pitch contour, glottal source and vocal tract (formants and bandwidths). The parameters are transformed only for voiced regions of speech.

As shown in [6] segmentation of the utterance is needed due to speech variability. So we split a utterance into segments, performing voice conversion for each segment. The results for this approach work well, giving good quality and the characteristics of the target speaker are reasonably well matched.

5. Voice Conversion at the Synthesizer

To add a new voice to the synthesizer we follow these steps:

- First we collect the 455 units from the new speaker.
- We apply the voice conversion algorithm to every diphone, and generate the coefficients to convert the unit from the basic speaker into the unit from the new speaker.
- If we choose the new speaker when synthesizing, voice conversion is performed before prosodic modifications and the new voice is generated.

Although the voice conversion algorithm converts the gain contour, in the synthesizer there is not gain control. We do not include gain conversion in the synthesis step.

5.1. Pitch conversion

In the synthesizer, the prosody of the original unit is modified to fit the prosodic model. It has no sense to apply the pitch conversion function to each segment in each unit because the prosody of the recorded diphones has nothing to do with the desired one.

We calculate a new transformation function for the speaker, that is, with the mean and variance of the fundamental frequency of each speaker we obtain the coefficients A,B of the linear function. This will be the transformation applied to the output of the prosody generator and will be constant for each speaker.

5.2. Transformation of formants, bandwidths and source parameters

For all these parameters voice conversion is done segment by segment. Formants are smoothed at segment boundaries as well as at unit changes.

Figure 3 shows the results of the transformation for the formant trajectories.

6. Evaluation and results

We have analyzed two male voices and one female voice. Pitch and formants were manually revised for all of them and some adjustments had to be made.

Two tokens were employed in informal tests: *"la bodega del avión"* and *"mi mamá me mima"*. All the sounds are voiced because we wanted to test the quality of the converted voice and the transformation algorithm is applied only to voiced segments.

All the speakers give very good results when used as basic speaker. The quality of the synthesizer is comparable to the quality of LP-PSOLA. When using the voice conversion algorithm the results are also good and the speech represents well enough the characteristics of the target speaker.

Some other tests have been performed in order to assess the conversion capability. All the possible transformations have been performed for the utterance *"la bodega del avión"* and formants have been extracted at synthesis time. Results are shown in table 2. The numbers in bold-face are the mean of each formant for each speaker. The rest of the numbers represent the mean error between the formant i of the basic speaker and formant i of the desired speaker in absolute value and percentage of the mean formant value.

Results show that higher formants are better converted than lower ones, this could be because higher formants are more stable during the utterance. We think that the main errors come from unit transitions, the weakest point of the system. We should improve the smoothing algorithm to reduce the error and possible artifacts.

			Basic speaker		
			MALE1	MALE2	FEM1
Desired speaker	MALE1	F1	**399**	53 13%	64 16%
		F2	**1451**	153 11%	119 8%
		F3	**2448**	199 8%	170 7%
		F4	**3707**	202 5%	200 5%
		F5	**5511**	252 5%	678 12%
	MALE2	F1	43 9%	**499**	105 21%
		F2	140 9%	**1592**	246 15%
		F3	148 6%	**2505**	163 7%
		F4	158 5%	**3350**	162 5%
		F5	285 7%	**4193**	388 9%
	FEM1	F1	72 14%	84 17%	**507**
		F2	163 9%	201 11%	**1825**
		F3	170 6%	194 7%	**2898**
		F4	178 4%	239 6%	**4135**
		F5	230 4%	299 5%	**5449**

Table 2

7. Conclusions and future work

We have developed a multi-speaker synthesizer by applying voice conversion techniques. The quality of the voice is equivalent to the results of an LP-PSOLA synthesizer.

A function as simple as a line is enough to transform the speaker when considering spectral stable segments of speech. Special attention must be paid to the transitions in order to avoid artifacts.

The system can be used to test the capability of isolated parameters when converting voices and transforming just the desired parameters.

One disadvantage of the proposed system is that we need all the data to apply the voice conversion algorithm. We are now working on extracting knowledge from the transformations to generate rules of conversion. We are also working on reducing the amount of data needed to extract the conversion functions.

8. References

[1] J.M. Montero, J.M. Gutiérrez-Arriola, S. Palazuelos, S. Aguilera, J.M. Pardo. “Emotional Speech Synthesis: from Speech Database to TTS”. Proc. ICSLP'98, vol. 3, pp 923-926. Sidney, 1998

[2] J.M. Montero, J.M. Gutiérrez-Arriola, J. Colás, E. Enríquez, J.M. Pardo. “Analysis and Modelling of Emotional Speech in Spanish”. Proc. 14th International Congress of Phonetic Sciences, vol. 2, pp 957-960. San Francisco, 1999

[3] H. Kuwabara, Y. Sagisaka. “Acoustic Characteristics of Speaker Individuality: Control and Conversion”. Speech Communication, vol. 16, pp 165-173. 1995

[4] L.M. Arslan. "Speaker Transformation Algorithm using Segmental Codebooks (STASC)" Speech Communication, vol.28, pp 211-226. 1999

[5] G. Fant, J. Liljencrants and Q. Lin, “A four-parameter model of glottal flow”. STL-QPSR, No 4/1985, pp 1-13.1985

[6] J.M. Gutiérrez-Arriola, Y.S. Hsiao, J.M. Montero, J.M. Pardo, D.G. Childers. “Voice Conversion Based on Parameter Transformation”. Proc. ICSLP'98, vol. 3, pp 987-990. Sidney, 1998

[7] F.M. Giménez de los Galanes, M.H. Savoji, J.M. Pardo. “Marcador automático de excitación glotal”. *Proc. URSI 93: 189-193. Valencia*.

[8] D.G. Childers, C. Ahn. “Modeling the glottal volume-velocity waveform for three voice types”. J. Acoust. Soc. Amer., vol. 97 (1), pp 505-519. 1995

[9] D.G. Childers, H.T. Hu. “Speech Synthesis by glottal excited linear prediction”. J. Acoust. Soc. Amer., vol.96, pp 2026-2036. 1994

[10] J.M. Gutiérrez-Arriola, F.M. Giménez de los Galanes, M.H. Savoji. “Improvement of the Quality of Speech Synthesis by Analysis Using Segmentation and Modelling of the Excitation Signal”. Proc. Eurospeech'95, vol. 2, pp 1097-1100. Madrid, 1995

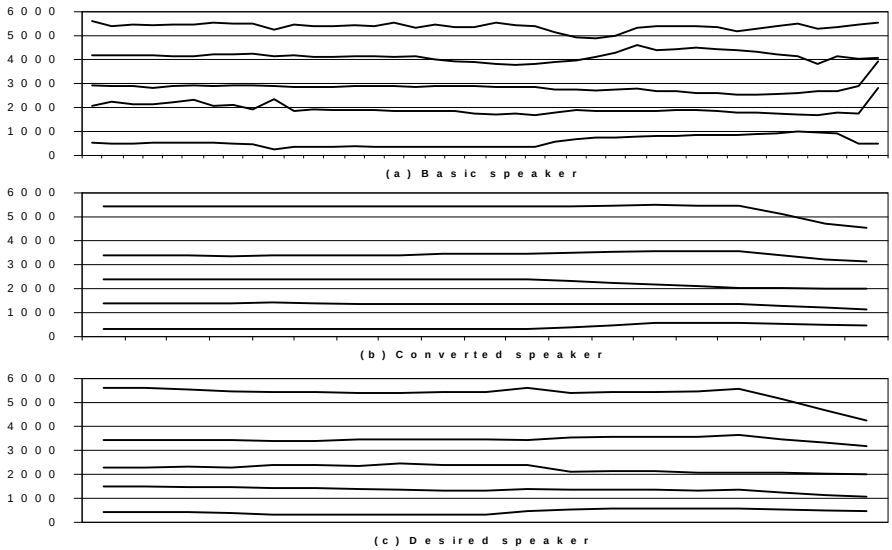

Figure 3: Formant conversion