ON THE CONSTRUCTION OF A PITCH CONVERSION SYSTEM

Tim Ceyssens, Werner Verhelst and Patrick Wambacq
Center for Processing Speech and Images (PSI)
Dept. of Electrical Engineering – ESAT
Katholieke Universiteit Leuven, BELGIUM
email :{tceys,wverhels,wambacq}@esat.kuleuven.ac.be*

ABSTRACT

In order to fully transform the perceived speaker identity, a voice conversion system should also convert the speaker's prosodic characteristics. When considering pitch contours, most systems only transform the pitch by simple scaling. A stochastic system that transforms pitch contours taking into account multiple pitch parameters, instead of only applying simple scaling, has been developed and will be described. A pitch transplantation system based on the overlap-add (OLA) algorithm is proposed as a tool for the evaluation of this pitch conversion system.

1 Introduction

One of the ultimate goals in speech modification research is the implementation of an automatic voice conversion system. Such a system takes an utterance of one speaker (the source speaker) and transforms it as if another speaker (the target speaker) had spoken it, and this with transparent quality, i.e., natural-sounding and without distortion. Speech modification being a relatively new domain, it is clear that nowadays systems are still far from this goal. High quality voice conversion systems have several applications [1]. As an example, text to speech synthesis systems could benefit greatly from it since the user would be able to choose a voice that suits him/her.

As there are several characteristics in a voice, it seems like a good strategy to begin with transforming each of these separately. In section 2 of this paper, a new stochastic system that transforms the pitch contour such that it resembles the one of the target speaker is presented.

As for the evaluation, it would be preferable that all voice characteristics, except the one being investigated in the conversion system, are transformed perfectly to the target utterance. This is however impossible for the obvious reason that such perfect transformations do not exist. The origin and the nature of this problem are explained in section 3, while section 4 describes the method used to overcome it. Finally, a concluding discussion is given in section 5.

*Tim Ceyssens is supported by a scholarship (Aspirant) of FWO-Flanders. Werner Verhelst is also with the Vrije Universiteit Brussel, Belgium.

2 Pitch Conversion System

2.1 Concept

It is obvious that conversion of timbre, i.e., how the *voice itself sounds*, is a main aspect of voice conversion. This would include modifying vocal tract parameters and voice source signal properties. As mentioned, the conversion of prosody should however also be incorporated into any voice conversion system. Parameters like speaking style and intonation do indeed contribute to the perception of speaker identity. Hence, an integrated approach is desired in which the conversion of certain intonation parameters also takes place, namely the pitch and the timing. Higher level information like the speaker's vocabulary and his strategy of stress placement also contribute to the perceived speaker identity. In this paper we study pitch conversion at the lower level, i.e., how conversion of the *way of stress production* can be done. This eventually results in a system that stresses the same syllables in the transformed utterance as in the source utterance.

Although pitch conversion systems that are more advanced in comparison to simple scaling have been studied before [2], many of the current voice conversion systems apply a constant scale factor to the pitch of the source speaker. This can be seen as a deterministic method. It would probably be better to view the pitch as a normal distributed entity and to modify mean and variance independently from one another. Furthermore, it is known that in a declarative phrase the pitch decreases overall. This is referred to as the pitch declination. Hence it is more accurate to determine declination lines of both speakers and to model the variations of the pitch around that line as normally distributed deviations. During the voice conversion one would then transform *both* the declination line and the deviations around it independently. This way an increasingly advanced transformation model can be built gradually. In a next step the *shape* and size of the deviations could

be modeled etc.

This concept consists of modeling the basic pitch parameters (mean, declination) in a *deterministic* way and the residual parameters in a *stochastic* way. The pitch conversion system described in this section transforms the declination line in the deterministic way and the residual in the stochastic way by simple-scaling it to the variance of the target's residual. In a next step, individual large pitch movements in this residual could be modeled, and the deterministic part of the model would increase while the residual's importance (magnitude) decreases. To obtain a perfect pitch conversion, this should be continued until the residual no longer has any perceptual value.

2.2 Training

In the training phase, the pitch contours of multiple utterances of all speakers of interest are processed in order to create a speaker-specific set of parameters, i.e., a set that characterizes the pitch behavior of the speaker.

The training procedure starts by transforming all pitch values to the log domain according to

$$P_{log} = 24\, log_2(P/110), \tag{1}$$

in which P is the pitch in Hz. The "reference" pitch of 110 Hz converts to 0 in the log domain, and 1 in the log domain is one semitone higher than the 110 Hz reference.

Next a regression line, calculated using the minimal quadratic error criterion, is added to the pitch contour plot of each utterance. Figure 1 shows this for one utterance. The offset (the intersection with the y-axis) and the slope of the regression line are determined. In the case of pitch these are referred to as pitch offset and pitch declination, Po and Pd, respectively. Then, the regression line is subtracted from the actual contour itself and the root-mean-square value of the resulting *residual* is calculated. This will be referred to as the variance V (even though it actually is a standard deviation). These steps are done for every utterance, each utterance giving 3 parameters. In the further processing only these 3 parameters and the length of all utterances are used.

Next, to model the dependence of these parameters on utterance length, all calculated values of each of these utterance parameters are plotted on y-axis with utterance length on x-axis and again a regression line is drawn. Correlations between the parameters on one side and the length on the other side will be accounted for this way. Figure 2 shows the correlation between pitch declination and utterance length. The two regression parameters (offset and slope) are used to model each of the parameters (Po, Pd, V) as functions of utterance length. For reasons to be made clear in the transformation procedure, the variance of the residual (regression line subtracted from parameter plot) is again calculated here.

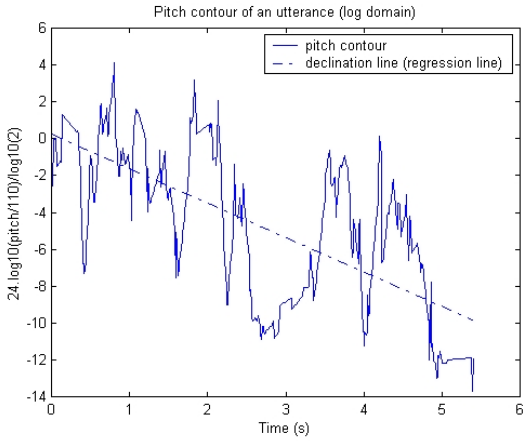

Figure 1: Pitch contour on a semi-tone scale (110 Hz reference) and declination line

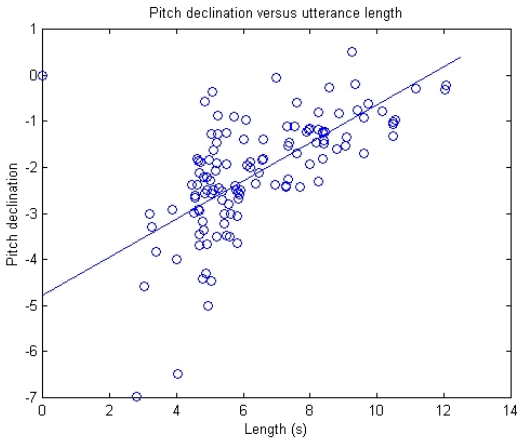

Figure 2: Pitch declination (semitones/s) versus utterance length

Eventually, this results in 9 parameters for each speaker, being offset, slope and variance for each of Po, Pd and V. They will be referred to as PSo_o, PSo_s, PSo_v, PSd_o, PSd_s, PSd_v, VS_o, VS_s and VS_v for the Source speaker and PTo_o, PTo_s, PTo_v, PTd_o, PTd_s, PTd_v, VT_o, VT_s and VT_v for the Target speaker. The "s" indexes stand for slope of regression lines, the "o" indexes for offset, and the "v" indexes for variance around the regression lines.

2.3 Transformation

When a source utterance is to be transformed, its original pitch contour will be given new pitch offset PCo, new declination PCd, and new variance VC (capital C for "Converted"). PCo will be calculated according to

$$PCo = (PTo_o + L.PTo_d) + PTo_v \frac{Po - (PSo_o + L.PSo_d)}{PSo_v}, \tag{2}$$

with L the length of the utterance and Po the pitch offset of the source utterance. PCd and VC are calculated in a similar way. This calculation takes into account the influence of the utterance length and also the locations of the source parameters in the source distributions. At this point, it is clear why the variances were calculated in the training phase. If, for example, the source utterance has a greater offset than the source's mean offset (given length L), the transformed utterance will also be given a greater offset than the target's mean offset.

Now that the transformation of the contour has been completed, it's just a matter of modifying the original utterance itself by a pitch shifting algorithm, such that the original utterance receives the desired calculated pitch contour. This step is included in the actions of the transplantation system described in section 4.

3 Evaluation Strategy

To do an evaluation, first thing needed is a database including several speakers. In the training procedure described in section 2.2, we just used a lot of utterances of each speaker to build a *parameter set* for each of them. In our evaluation itself it is however necessary to at least have a subset of *common* utterances, i.e., the same sentences spoken by all speakers, since we compare transformed utterances with the *same* utterance of the target speaker. The set of utterances on which training is done may be larger than the set used in the evaluation, i.e., the set of common utterances. In our training, we used a subset of the WSJ database.

An evaluation experiment in which the test person can use a window with multiple buttons was constructed. This window is depicted in figure 3. On the window the test person has three buttons that play utterances through a soundcard. Two of them are transformed versions of a source utterance. The first one is pitch-modified only by simple scaling and the other one is transformed by the previously described conversion system. Which button corresponds to which of the two transformations is unknown to the test person, since this is varied randomly. This results in a *blind* test. The third "play" button plays the same utterance, not transformed, spoken by the target speaker. The other buttons in the figure are self-explanatory. The test person can hit the three "play" buttons in any order and as many times as he/she wishes. Eventually the person has to decide which of the two transformed versions resembles the target utterance best. The evaluation procedure then just continues by going to the next utterance, and at the end one can see whether the new pitch conversion system was more successful than the reference system.

When evaluating in this straightforward way however, the test person is faced with a difficult situation. The problem is that both the transformed utterances still resemble the original (source) voice (S) much more than the target voice (T), because pitch is only one aspect

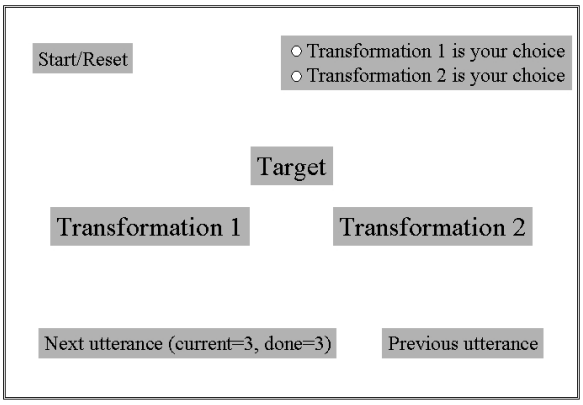

Figure 3: User window for evaluation

of the difference between the two voices. This makes it difficult for the test person to decide. Therefore it would be better if all the remaining voice characteristics would be transformed *perfectly* to the target. This is however impossible for the obvious reason that such perfect transformations do not exist. Therefore it would be convenient to use a system that is able to construct an utterance by extracting each voice characteristic from one of two given utterances. This system would then be applied to synthesize test utterances UX, extracting all of the voice characteristics from the target T except the one being investigated, in this case pitch. The latter would be extracted from the source S. Transforming by simple scaling of the pitch results in what will be referred to as $UX1$, while transforming by applying the described conversion system results in $UX2$. This way, the test person has an easier task choosing between $UX1$ and $UX2$ when comparing them with T. An evaluation tool capable of this was constructed and will be described in the next section. At the time of writing, experiments using this tool are being planned in order to evaluate the pitch conversion system of section 2.

4 Pitch Transplantation

The system architecture started from is a prosodic transplantation system which has been described earlier. For a thorough explanation of it we refer to the literature [3, 4]. It deals with a generic system in the sense that it can synthesize by first extracting pitch p, windowed waveform h and loudness α contours from two utterances during analysis, then using either one of each pair during synthesis. For the timing, it can also choose between the two utterances during synthesis. The prosodic characteristics here are p, α and the timing, and all of these can be transplanted, hence the name prosodic transplantation system.

In the evaluation of the described pitch conversion system, only the transplantation of pitch was needed.

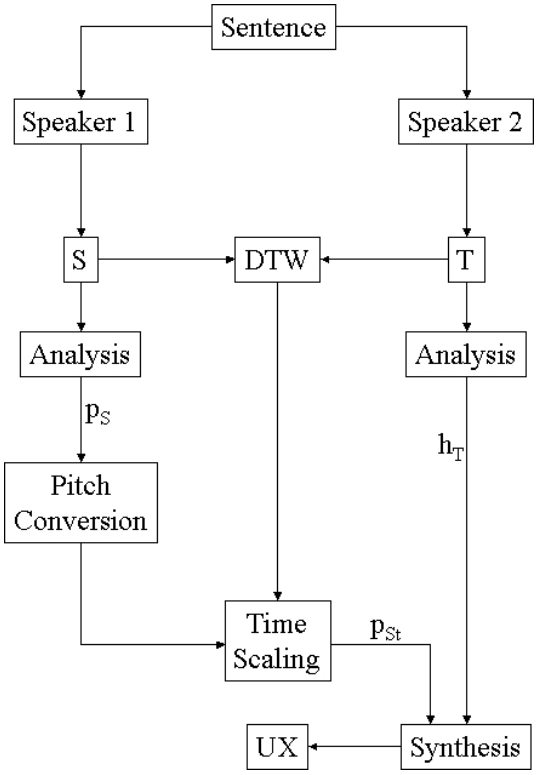

Figure 4: Pitch transplantation system

Therefore, the transplantation system was not needed in its generic form. Instead, a simplified form was implemented of which the block diagram is depicted in figure 4. A sentence is spoken by a source and a target speaker, resulting in utterances S and T respectively. Both S and T are subjected to PIOLA analysis [5, 6]. The pitch of S is measured using a pitch detection algorithm (PDA). This results in p_S. The other characteristics are obtained from the analysis of T resulting in h_T (the influence of α_T is included in h_T here). Apart from these analyses, a dynamic time warping algorithm (DTW) is used to determine the time alignment path between the two utterances S and T. After conversion, the source's pitch contour is time-scaled to the timing of the target T. This results in p_{St}. During synthesis, p_{St} and h_T are combined to construct the test utterance UX.

5 Concluding Discussion

We presented a methodology for the construction of a pitch conversion system for voice transformation applications. This methodology is targeted at automatically learning the transformation rules for the manner in which speakers use pitch to produce a given stress pattern.

The central element in our methodology is a hybrid deterministic/stochastic modeling strategy for pitch contours. This strategy allows for gradual increments in the amount of knowledge that is deterministically modeled (the residual difference with the actual contour being modeled stochastically). Basically, the deterministic part of the model approximates the pitch contour by straight lines on a logarithmic scale, as is the case in many pitch models, and the deterministic approximation error is modeled as Gaussian noise.

The traditional strategy of converting the mean and variance of a speaker's pitch values can thus be understood as a zero-order version of the proposed approach. In this paper, we introduced the first-order version, in which the deterministic component is the declination line of the pitch contour.

In order to evaluate the quality of our first-order pitch conversion model, a pitch transplantation system was conceived. With this, it becomes possible to transplant a converted pitch contour onto the target speaker's version of the same utterance. This way a voice conversion system is simulated that uses the proposed pitch conversion rule and performs a perfect copy conversion of the remaining speech characteristics.

At the time of writing, experiments are being planned to evaluate to what extent our first-order pitch conversion could improve over the traditional mean-and-variance (i.e., zero-order) approach. If the quality of the result turns out not high enough in absolute terms, our hybrid modeling strategy allows for a gradual increase in the amount of explicitly (deterministically) modeled detail. For example, a second-order model could include a straight-line approximation of the accent lending pitch movements in the deterministic part of the model.

References

[1] P.F. Yang and Y. Stylianou, "Real time voice alteration based on linear prediction," in *Proc. ICSLP*, Sydney, Australia, Nov. 1998, pp. 1667–1670.

[2] D. Chappell and J. Hansen, "Speaker-specific pitch contour modeling and modification," in *Proc. ICASSP*, Seattle, U.S.A, May 1998, pp. 885–888.

[3] W. Verhelst and M. Borger, "Intra-speaker transplantation of speech characteristics," in *Proc. EUROSPEECH*, Genova, Italy, Sept. 1991, pp. 1319–1322.

[4] W. Verhelst, "Automatic postsynchronization of speech utterances," in *Proc. EUROSPEECH*, Rhodes, Greece, Sept. 1997, pp. 899–902.

[5] W. Verhelst, D. Van Compernolle, and P. Wambacq, "A unified view on synchronized overlap-add methods for prosodic modification of speech," in *Proc. ICSLP*, Beijing, China, Oct. 2000, vol. II, pp. 63–66.

[6] L.L.M. Vogten, C. Ma, W.D.E. Verhelst, and J.H. Eggen, "Pitch inflected overlap and add speech manipulation," European patent 91202044.3, granted to Philips nv., 1991.