Transforming F0 Contours

Ben Gillett and Simon King

Centre for Speech Technology
University of Edinburgh, United Kingdom
beng@cstr.ed.ac.uk, Simon.King@ed.ac.uk

Abstract

Voice transformation is the process of transforming the characteristics of speech uttered by a source speaker, such that a listener would believe the speech was uttered by a target speaker. Training F0 contour generation models for speech synthesis requires a large corpus of speech. If it were possible to adapt the F0 contour of one speaker to sound like that of another speaker, using a small, easily obtainable parameter set, this would be extremely valuable. We present a new method for the transformation of F0 contours from one speaker to another based on a small linguistically motivated parameter set. The system performs a piecewise linear mapping using these parameters. A perceptual experiment clearly demonstrates that the presented system is at least as good as an existing technique for all speaker pairs, and that in many cases it is much better and almost as good as using the target F0 contour.

1. Introduction

Voice transformation also has other applications such as very low bandwidth speech encoding, multimedia entertainment, as a pre-processing step to speech recognition and also in the field of voice disguise. In addition, gaining a better understanding of the ways in which speakers differ is likely to be valuable more generally in both speech synthesis and recognition. Training F0 contour generation models for speech synthesis requires a large corpus of speech [1].

Very little work has been directed at the problem of mapping the F0 contours of one speaker to another. The approach taken by existing systems [2] is to simply normalise the F0 of the source speaker to be like that of the target speakers. We will call this mapping function M_N, where

$$M_N(x) = ((x - \mu_{\text{src}})/\sigma_{\text{src}}) * \sigma_{\text{targ}} + \mu_{\text{targ}} \qquad (1)$$

and μ_{src}, σ_{src} are the mean and standard deviation of the source speaker respectively, and μ_{targ}, σ_{targ} are the mean and standard deviation of the target speaker.

This mapping technique fails to capture many of the important features of F0 contours, which contain information about speaker identity. We present a method for the transformation of F0 contours from one speaker to another based on a small linguistically motivated parameter set. This was first presented in an earlier paper [3].

2. Parameterisation

We use the parameterisation described by Patterson [4], which was based on work by Ladd and Terken [5]. Patterson took F0 measurements at four selected target points in each sentence. These points were sentence-initial high (S), non-initial accent peaks (H), post-accent valleys (L), and sentence-final low (F). For each sentence there is one sentence-initial high, one sentence-final low and a varying number of peaks and valleys depending on the sentence. Patterson carried out analysis on approximately a minute of speech for each speaker. The values were collected into their respective categories and then averaged to get representative data for the speaker. Figure 1 shows diagramatically where the four points lie. The mean and standard deviation of the F0 of the voiced segments of speech for each speaker were also computed. In this work we make use of the values of S,H,L,F, mean and standard deviation collected by Patterson. The following work is entirely that of the authors and was not proposed by Patterson. We are simply using his data set and measurements of the parameters.

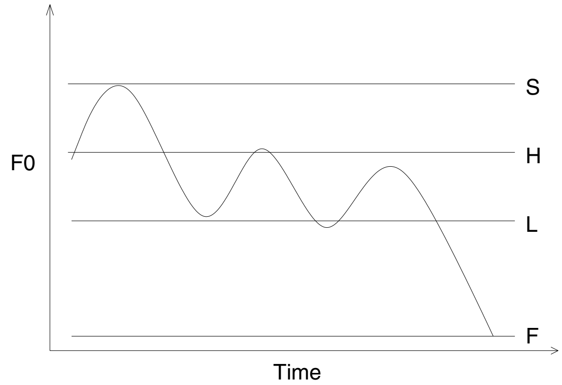

Figure 1: Measurement locations on an idealised speaker contour.

3. Mapping

The mapping from source to target F0 is then defined by a piecewise linear mapping, where one segment runs through the points $(F_{\text{src}}, F_{\text{targ}})$ and $(L_{\text{src}}, L_{\text{targ}})$, another between $(L_{\text{src}}, L_{\text{targ}})$ and $(H_{\text{src}}, H_{\text{targ}})$, and a final segment through $(H_{\text{src}}, H_{\text{targ}})$ and $(S_{\text{src}}, S_{\text{targ}})$. An example mapping is shown in figure 4, where one can see how a value x may be transformed to a value $M_{PL}(x)$. The mapping function M_{PL} is:

$$M_{PL}(x) = \begin{cases} F_{\text{targ}} + \frac{(x - F_{\text{src}})(L_{\text{targ}} - F_{\text{targ}})}{(L_{\text{src}} - F_{\text{src}})} & x < L_{\text{src}} \\ L_{\text{targ}} + \frac{(x - L_{\text{src}})(H_{\text{targ}} - L_{\text{targ}})}{(H_{\text{src}} - L_{\text{src}})} & L_{\text{src}} \leq x \leq H_{\text{src}} \\ H_{\text{targ}} + \frac{(x - H_{\text{src}})(S_{\text{targ}} - H_{\text{targ}})}{(S_{\text{src}} - H_{\text{src}})} & x > H_{\text{src}} \end{cases} \qquad (2)$$

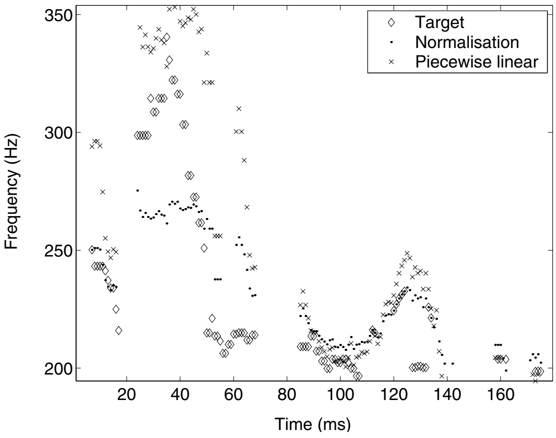

Figure 2: Target and mapped F0 tracks (S:fl T:so)

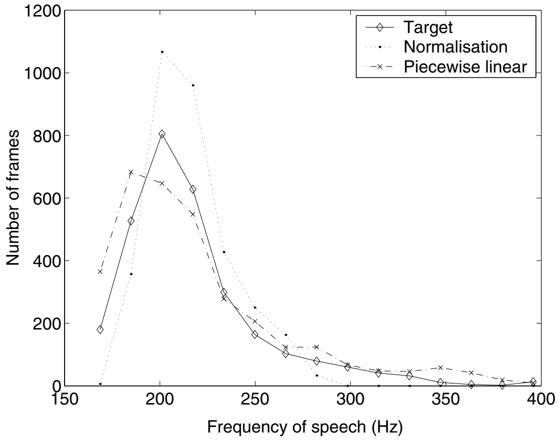

Figure 3: Histogram of frequencies for one minute of speech from the Patterson corpus (S:fl T:so)

4. Transformation

Pitchmarks and F0 tracks are first found for the speech to be transformed, using the ‘pitchmark’ and ‘pda’ programs from Edinburgh Speech Tools [6]. The four parameters (S,H,L,F) were obtained for both source and target speaker (from [4]). These eight parameters were then used to define the mapping M_{PL}. For each voiced frame of the F0 track, the F0 value was converted using M_{PL}. Finally, pitchmarks were generated from the transformed F0 track, and the speech was resynthesised using pitch synchronous overlap and add (PSOLA) [7]. Example transformed speech is available online [8]. Figure 2 shows that the F0 mapped using M_{PL} more closely follows the target contour than that mapped using M_N. Figure 3 shows that the distribution of F0 frequencies in one minute of speech uttered by speaker ’so’ is more accurately captured using M_{PL} than M_N.

5. Evaluating the F0 transformation system

We wish to ascertain if the proposed method is perceived as producing contours that are more similar to the F0 contours of the target speaker, than the existing technique. In order to do this, we conducted a perceptual experiment.

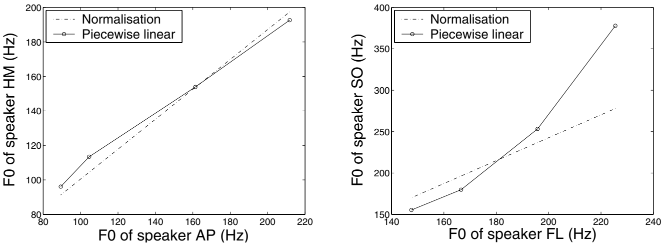

Figure 4: Graphs showing an example of a frequency mapping where the piecewise linear mapping is very similar to the normalisation mapping (left), and a frequency mapping where they are very different (right).

5.1. Measuring the difference between techniques for given speaker pairs

In this experiment we will be investigating the relative effectiveness of the new method we presented in the last chapter (M_{PL}), against the existing technique based on the normalisation of mean and standard deviation (M_N). The extent to which the results of the two methods differ is dependent on the particular parameters of the two speakers involved. For example, if the four points of the piecewise linear mapping $((F_{\text{src}}, F_{\text{targ}})$, $(L_{\text{src}}, L_{\text{targ}})$, $(L_{\text{src}}, L_{\text{targ}})$, $(H_{\text{src}}, H_{\text{targ}}))$, lie on the line defined by the mean and standard deviations $(\mu_{\text{src}}, \sigma_{\text{src}}, \mu_{\text{targ}}, \sigma_{\text{targ}})$, then the result of applying these two mappings will be identical. Figure 4 shows one mapping where the difference is large, and another where the difference is small.

The extent of any preference for one or other technique is likely to be proportional to the degree by which the two techniques differ for the speaker pair being tested. Therefore we have devised a method for determining how different the mappings are, for a particular speaker pair.

The difference between the two techniques for a given speaker pair is computed by taking the sum of the squares of the differences between the mapped frequencies generated by each of the two methods, at points corresponding to $S_{\text{src}}, H_{\text{src}}, L_{\text{src}}, F_{\text{src}}$. This difference can be represented as:

$$\begin{aligned} D'(A,B) &= (M_N(S_{\text{src}}) - S_{\text{targ}})^2 + (M_N(H_{\text{src}}) - H_{\text{targ}})^2 \\ &\quad + (M_N(L_{\text{src}}) - L_{\text{targ}})^2 + (M_N(F_{\text{src}}) - F_{\text{targ}})^2 \end{aligned}$$

where all frequencies are measured on an equivalent rectangular bandwidth (ERB) scale [7].

This difference measure is not symmetric. In other words, $D'(A,B) \neq D'(B,A)$. This is as expected, since the mapping function defined in the last chapter is non-linear. However, it is likely that there will be a high correlation between the two values. It is useful to have an overall distance measure between two speakers. We define this to be:

$$D(A,B) = \frac{D'(A,B) + D'(B,A)}{2} \tag{3}$$

5.2. Stimuli

5.2.1. Generation of stimuli

The speech used in this experiment was recorded previously by Patterson [4]. Two sentences were selected from this corpus, chosen for their relatively short duration.

> 1) 'Madonna has been lined up as a key backer along with Ossie Kilkenny, the accountant to the stars.'
> 2) 'Kilkenny, whose clients include the rock band U2, will be employed as a consultant.'

Seven male and seven female speakers were selected, all of whom are native speakers of English, and have an accent commonly spoken by people from the Home Counties in England. Their ages range from 19 to 65.

For each same sex speaker pair (S, T), and for each of the two sentences, we created three stimuli. Firstly, the sentence uttered by source speaker S with its F0 modified to have the mean and standard deviation of the target speaker T. Secondly, the source speech with its F0 modified using the new method presented in this paper. Finally, the source speech with the actual F0 contour of the target applied to it. This final sentence is the ideal output of an F0 transformation system. For all three stimuli types, although the F0 was modified, the voice quality was not. Therefore these stimuli have the voice quality of the source speaker with the intonation of the target speaker.

The experiment was of an XABX type, where X was the sentence spoken by the target speaker. A and B were the same sentence spoken by the source speaker modified to have an F0 contour like the target speaker, by one of the three methods described earlier. The decision to make the experiment XABX rather than ABX was based on the fact the utterances are relatively long, and in pilot experiments it was found that playing the target twice helped the subjects decide which of A or B was better.

5.2.2. *Grouping of stimuli*

In order to ensure that any result we obtain is due to improvements in our method over the existing technique, we created two sets of stimuli. One of these groups consisted of those where we expect to get a clear preference for one or other method. These were the speaker pairs where the distance between the two methods, as defined in equation 3, was large. We call this group $S_{\text{different}}$. The other group was of speaker pairs where the two methods do not differ greatly. We call this group S_{same}. For each sex we selected five pairs where $D(S, T)$ is large, and five where it is small. We did not necessarily select the five largest or smallest values, since we were also trying to ensure that each speaker is chosen about the same number of times.

In order to control for ordering effects, any given pair of stimuli was always presented both ways round to a given subject. Since there are three methods (A,B,C), there are 6 possible combinations as follows: XABX, XBAX, XCAX, XACX, XBCX, XCBX.

Since the concentration span of our subjects is limited, and we are most interested in the result concerning the destinction between the existing method (A), and our newly presented method (B), each subject is either presented with (XABX, XBAX, XCAX, XACX) or (XABX, XBAX, XCBX, XBCX). This is carried out such that there are always paired groups, so that all those subjects with odd subject numbers are presented with the first set, and all those with even numbers with the second set.

Since there are two groups of stimuli (one for each sex), and for each of these there are five pairs of speakers which must be presented both ways round, and there are two sentences, and four method combinations as described in the last paragraph, there are $2 * 10 * 2 * 2 * 4 = 320$ trials to be run. However, since each trial takes approximatly 30s, and we wish to restrict the experiment to not more than 35 minutes so that concentration is not impaired, we can have at most 70 trials per subject. Therefore, for each subject we select four speaker pairs, where half are from S_{same} and half from $S_{\text{different}}$, to give a total of 64 stimuli for each subject.

5.3. Subjects

Twenty-five subjects were selected of whom approximately half were native and half non-native speakers of English. Similarly, approximately half were male and half were female. The task of discriminating between two similar F0 contours is difficult. Therefore a relatively large number of subjects where selected, with a view to removing those who were not good at the task from the analysis.

5.4. Experiment

The E-Prime experiment design system was used for this experiment [9]. The subjects were placed in a quiet booth with headphones, computer monitor and an input box for recording responses. The subjects were given on-screen instructions regarding the procedure for inputing data. The following instructions were then given to the subjects:

> You will be presented with four pieces of speech. First you will be presented with a piece of target speech. Then two attempts of a speaker at imitating the F0 of the target speaker. Finally you will be presented with the target speech once more.
>
> (Target) target speech
> (1) first attempt by the imitator
> (2) second attempt by the imitator
> (Target) target speech
>
> You must decide whether attempt 1 or 2 has a more similar pitch pattern to the target. You shouldn't make your decision based on any aspects of the voice apart from pitch.
>
> If you think the first attempt sounds most like the target then press 1. If you think the second attempt sounds most like the target then press 2.
>
> In some cases it will be very hard to distinguish a difference between attempts 1 and 2. If so, just choose one or other.
>
> If you have any questions, please ask the experimenter now, otherwise press either button to continue.

The subjects were given three practice trials, followed by 70 actual trials. The experiment took approximately 35 minutes for each subject to complete. The order in which the stimuli were presented was randomized for each subject.

6. Results

6.1. Rejecting poor subjects

For each subject, the number of times they selected the ideal contour in preference to a contour formed by either of the mapping techniques was counted. In order for the results to be meaningful, the subject must be capable of telling that the 'correct' F0 contour is better at representing the target speaker than a contour formed by either of the mapping techniques. Since the task is difficult, a relatively low level (60%) of preference for the ideal contour was selected as a criteria for rejecting subjects

	Mean (%)	Std. Dev.	α	t
Preference for M_{PL} over M_N for $S_{\text{different}}$	67	10	$< 1 \times 10^{-7}$	-8.71
Preference for M_{PL} over M_N for S_{same}	54	8	~ 0.02	-2.49
Preference for target over mapped contours	73	9	$< 1 \times 10^{-11}$	-13.8

Table 1: Table showing the subject preferences for different mapping methods.

from further analysis. Just over half (13 of 25) of the subjects were able to tell that the ideal contour was better than mapped contours. It may seem suprising that such a high proportion of the subjects were not able to distinguish effectively, however making such judgements is difficult for naive listeners.

6.2. Statistical analysis

For the remaining set of subjects, the number of times the subject preferred contours mapped with M_{PL} over M_N was counted for each of the two data sets, $S_{\text{different}}$ and S_{same}. No correlation between the nature of the subject (i.e. whether they were native speakers, their sex), and their preferences was found. Also, the sex of the speaker does not appear to make any difference to the preferences expressed.

The mean and standard deviation of each category was then computed. In order to establish the statistical significance of these results, we used Student's t-test for equal variances. A one tailed analysis was performed, since we are trying to determine the probability of a particular method being better than the other, rather than looking for a preference either way. The results of this analysis are contained in table 1. Students t-test provides a value of α, where α indicates the probability of the result being purely due to chance. A value of $\alpha \leq 0.01$ is generally accepted as being a statistically significant result. It is therefore clear from table 1 that the preference for M_{PL} over M_N for $S_{\text{different}}$ is highly significant. Similarly the preference for the target contour over the mapped contours is also significant. However, the significance of the preference for M_{PL} over M_N for S_{same} is not very high, as is to be expected, since on the data set S_{same}, the two methods (M_{PL} and M_N), are almost identical (see section 5.1).

7. Conclusions

It was found that 73% of the time subjects expressed a preference for the ideal contour over a mapped contour. The remaining 27% of the time the subject chose the mapped contour, this is likely to be due to the fact that the contours mapped with the two methods were so similar that the subject was not able to distinguish between them. A clear preference for our method is shown in the experiment, with subjects selecting the speech modified with the mapping M_{PL} in preference to M_N for the dataset where the two methods are most different 67% of the time. This result compares very favourably with the preference for the ideal contour of 73%, suggesting that using M_{PL} is almost as good as using the actual contour. In the cases where the mapping techniques differ least, there was a preference for M_{PL}, although it is on the border of not being statistically significant.

It has been clearly shown that the presented method based on a piecewise linear mapping is at least as good as the only existing technique for F0 contour mapping for *all* speaker pairs, and that in *many cases* it is much better and almost as good as using the target F0 contour.

The work on F0 transformation makes use of a number of parameters that were extracted by hand. However, for this approach to be useful, methods must be developed which extract these parameters automatically. The problem of finding these parameters is likely to be much easier than finding the pitch accents in a sentence. In order to find the sentence intial high, one may simply find the highest F0 in the first one second of speech, and similarly one may find the sentence final low by finding the minimum of the last second of speech for the sentence. To find sentence medial highs and lows, an approach based on finding maxima and minima in a smoothed F0 contour may well produce good results.

8. Acknowledgements

The authors are grateful to Bob Ladd for his guidance and assistance.

9. References

[1] Alan Black and Andrew Hunt, "Generating f0 contours from tobi labels using linear regression," in *Proceedings of the International Conference on Speech and Language Processing*, 1996.

[2] Levent M. Arslan and David Talkin, "Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum," in *Proc. Eurospeech '97*, 1997, pp. 1347–1350.

[3] Ben Gillett, "Transforming voice quality and intonation," M.S. thesis, Centre for Speech Technology Research, University of Edinburgh, 2003.

[4] David Patterson, *A Linguistic Approach to Pitch Range Modelling*, Ph.D. thesis, University of Edinburgh, 2000.

[5] Bob Ladd and J. Terken, "Modelling intra- and inter-speaker pitch range variation," in *Proceedings of the International Conference of Phonetic Sciences*, 1995, vol. 2, pp. 386–389.

[6] Paul Taylor, Richard Caley, Alan W. Black, and Simon King, *Edinburgh Speech Tools Library : System Documentation*, Centre for Speech Technology, University of Edinburgh, 1999.

[7] Ben Gold and Nelson Morgan, *Speech and Audio Signal Processing*, Processing and Perception of Speech and Music. John Wiley and Sons, Inc., 2000.

[8] Ben Gillett, "Audio examples to accompany 'transforming voice quality and intonation'," http://www.cstr.ed.ac.uk/voicetransformation, 2003.

[9] Psychology Software Tools, Inc., *E-Prime Users Guide*, 2002, http://www.pstnet.com/e-prime/default.htm.