INTERSPEECH 2009 BRIGHTON

An Evaluation Methodology for Prosody Transformation Systems based on Chirp Signals

Damien Lolive, Nelly Barbot, Olivier Boeffard

IRISA / University of Rennes 1 - ENSSAT
6 rue de Kerampont, B.P. 80518, F-22305 Lannion Cedex
France

{damien.lolive,nelly.barbot,olivier.boeffard}@irisa.fr

Abstract

Evaluation of prosody transformation systems is an important issue. First, the existing evaluation methodologies focus on parallel evaluation of systems and are not applicable to compare parallel and non-parallel systems. Secondly, these methodologies do not guarantee the independence from other features such as the segmental component. In particular, its influence cannot be neglected during evaluation and introduces a bias in the listening test. To answer these problems, we propose an evaluation methodology that depends only on the melody of the voice and that is applicable in a non-parallel context. Given a melodic contour, we propose to build an audio whistle from a chirp signal model. Experimental results show the efficiency of the proposed method concerning the discrimination of voices using only their melody information. An example of transformation function is also given and the results confirm the applicability of this methodology.

Index Terms: subjective evaluation, prosody transformation, chirp signal, non-parallel corpora

1. Introduction

These last years, voice transformation systems are becoming more and more used and studied as it is shown by an increasing number of papers in the literature. Applications can be found in fields like biometric identification or text-to-speech (TTS) systems. In the former field, a voice transformation methodology can be used to test a speaker identification or verification system, while in the latter field, a possible use is to facilitate the production of synthetic voices.

In this paper, we focus on prosody transformation systems. The goal of such a system is the transposition of a segmental transformation system to the supra-segmental level and it has to modify the prosody of a sentence pronounced by a source speaker so as to be perceived as if it were uttered by another speaker, the target one. In our case, we focus more particularly on the fundamental frequency, F_0, transformation.

Transformation systems commonly described in the literature are constrained to have parallel corpora, as in [1, 2]. Two corpora are considered parallel if they contain sentences uttered by the source and the target speakers that have the same linguistic content. This point greatly constrains the use of transformation systems since it implies the learning and testing stages use parallel corpora, and the size of the source corpus and the target one are identical.

Recently transformation methodologies have been proposed to circumvent this constraint and facilitate the development of such systems. For instance, we can cite the well-known Gaussian normalisation transformation function and the one described in [3] that uses a Maximum Likelihood Linear Regression (MLLR) based strategy to learn a transformation function. Such methodologies permit to build a transformation function which is non-parallel but do not provide an answer to the problem of subjective evaluation in this particular framework.

Indeed, the evaluation of the transformation function is a crucial aspect for every voice or prosody transformation system. In this field, even if there is no standard evaluation protocol, some techniques using objective or subjective criteria are commonly used. The former are composed of error measures used to evaluate the proximity between the transformed prosody and the target one while the latter commonly use ABX or MOS tests. Three main problems have to be treated:

- the evaluation in an independent manner of the transformation of the different prosodic features [4]. For example, the evaluation of a F_0 transformation function must not be disturbed by the segmental component.
- the evaluation of non-parallel transformation systems. Given that a transformation function is learnt in a non-parallel framework, the evaluation process has to be done coherently.
- the comparison of transformation systems in a common framework. Existing results in the field of voice transformation and in particular prosody transformation cannot be compared. A reason for this can be found in the fact that no common evaluation framework exists.

A possible answer to these problems is presented in this paper and is based on chirp signals to generate a waveform that depends only on the fundamental frequency contour and on ABX type listening tests that are built in a non-parallel manner.

In section 2, the existing evaluation methodologies are discussed. Next, in section 3, the proposed approach which is based on chirp signals is detailed. Then the experimental protocol is described in section 4 and the results are finally discussed in section 5.

2. Evaluation of prosody transformation

Two kinds of evaluation strategies can be considered: subjective and objective evaluations. The former is subjective in the sense that the result depends on the opinion and the expertise level of a group of testers. In the field of automatic speech processing, and in particular for a system that produces a speech signal, necessarily, the final user must evaluate the quality and

the natural of the produced voice. This evaluation can also be considered as objective when listener panels and a precise experimental methodology that guarantee the stochastic interpretation of the results are used.

An objective evaluation is based on a measure, most of the time an error measure or a model similarity measure, that must be representative of the goal of the evaluation. As an example, is a low quadratic mean error (RMSE, Root Mean Square Error) for the F_0 representative of the perceived quality of a F_0 transformation ? We can say that if that error is zero, then the transformed contour is perfect. However, what is the link between the evolution of that error and the perceived degradation ? These questions show the complementary of both objective and subjective approaches.

2.1. Subjective evaluation

The subjective evaluation of a transformation system must use at least two types of data which are transformed data and target speaker data. According to Ceyssens & al. [4], it is necessary to evaluate the transformation of only one supra-segmental acoustic characteristic at a time, the others are then just copied from the target speaker to the source speaker. This point of view brings the constraint of having parallel speech corpora of the source and target speakers. Nonetheless, this approach enables the evaluation of the contribution of each characteristic independently of the others. In the framework of parallel corpora, an ABX preference test, which consists in deciding if X resembles more A than B, can be made in the style of Ceyssens & al.[4]. The tester gives a binary answer that may be extended as it is done by Hanzlicek and Matousek [5], by considering a gradual answer with five levels (like A, rather A, no difference, rather B, like B). However, Helander and Nurminen [2] use an evaluation methodology that removes the constraint of parallel corpora. This test is used to determine which transformation method, among two, the user prefers. The experimental protocol is the following:

- the tester listens to a set of sentences spoken by the target speaker for him to familiarize with its elocution style,
- the test is realised by proposing, at each step, one sample from each method,
- the testers are asked to choose the sample whose prosody is closest to the prosody that the target speaker could use,
- test sentences are different from the sentences presented for the target speaker.

This methodology is similar to situations that occur in everyday life. As an example, when we receive a phone call, we recognize the person thanks to a certain habit of hearing his timbre as well as his elocution style. The previous methodology follows this observation and presents, to the tester, sentences pronounced by the target speaker first and then the test sentences.

These evaluations are used to assess the level of resemblance between the transformed prosody and the prosody that the target speaker could have used. We can also evaluate other aspects such as the quality and the natural of the transformed prosody by applying Mean Opinion Score tests (MOS). The goal of this type of test is to compare several systems by quoting their performance.

A subjective test is built with spoken messages. Therefore, it uses both acoustic segmental and supra-segmental characteristics at the same time. These characteristics are interdependent, and comparing the prosody of two voices that have a different segmental component is difficult. A manner to solve this problem, in the case common sentences are available for the test, is to "transplant" the unmodified supra-segmental characteristics from one speaker to the other, in order to keep the segmental component of only one speaker at each test step.

2.2. Parallel versus non parallel corpora

We have used several times the terms of parallel corpus and non parallel corpus. Let us recall that we assume in this article that two corpora are considered parallel when the sentences of the two corpora have the same linguistic content.

This strong constraint is often imposed by the models for which parameter learning needs a data parallelism between source and target speakers. For example, this is the case for the codebook-based transformation. A major drawback of such corpora is that their development cost is quite important. Relaxing this constraint would soften the conception of applications and it is an important issue for voice transformation systems and in particular for prosody ones.

3. Proposed approach

3.1. Principle

In the framework of non-parallel corpora, common evaluation methodologies are difficult to apply. Moreover, it is desirable to isolate the phenomenon under test. In our case, we focus on the evaluation of listener's ability to discriminate voices only by means of their melody.

The methodology we propose to achieve this goal relies on the use of sound samples built from F_0 contours of both speakers while excluding the influence of timber perception and the influence of the nature of sounds.

The stimuli created with this method are then used in a subjective evaluation that relies on non-parallel corpora and consists in two ABX tests. The evaluation of prosody transformation systems we propose is presented in paragraph 3.4.

3.2. Creation of stimuli

The procedure implemented to create the experimental stimuli is based on an original method that enables the generation of a sinusoidal signal by frequency modulation of the F_0 curve (chirp signal). This way, a sinusoidal signal is described by

$$s(t) \;=\; \sin(\psi(t)) \tag{1}$$

where $\psi(t)$ is the instantaneous phase of signal $s(t)$. The instantaneous frequency of signal $s(t)$ is defined by

$$f(t) \;=\; \frac{1}{2\pi}\frac{d\psi(t)}{dt}\,.$$

Our goal is to obtain a signal of which instantaneous frequency f is equal to melody function F_0 at any sampling instant t_i. We assume that the F_0 sampling period is a constant value, denoted by T, and f is a piecewise linear function that interpolates the point sequence $\{(t_i, F_0(t_i))\}$. Let us consider two consecutive points at instants t_i and t_{i+1}, and denote $\Delta F_i \;=\; F_0(t_{i+1}) - F_0(t_i)$. For any $t \in [t_i, t_{i+1})$ we have

$$f(t) \;=\; F_0(t_i) + \frac{\Delta F_i}{T}(t - t_i)$$

and then

$$\psi(t) \;=\; \psi_i + 2\pi\left(F_0(t_i)(t - t_i) + \frac{\Delta F_i}{2T}(t - t_i)^2\right) \tag{2}$$

where (ψ_i) is chosen such as phase ψ is a continuous function at every instant t_i in order to insure the absence of audible artefacts in the modulated signal. Therefore, (ψ_i) is defined by $\psi_0 = 0$ and, for $i \in \mathbb{N}$,

$$\psi_{i+1} \; = \; \psi_i + \pi T \left(F_0(t_{i+1}) + F_0(t_i) \right) \; .$$

3.3. Example of stimuli

An example of the use of this methodology is given in Fig. 1a and 1b. The first part of the figure shows the orginal waveform uttered by the speaker and corresponding to the french sentence "mais tout le monde dans l'ancien Far West ne menait pas une vie aventureuse !". The associated F_0 contour is drawn below. Applying the above methodology, this F_0 contour is used to generate a waveform (shown in 1b) whose the instantaneous frequency is equal to F_0 at each F_0 sampling instant. The generated signal is given by equation (1) and has a continuous phase defined by equation (2). We can observe that the F_0 contour estimated from the generated signal is nearly identical to the F_0 contour of the original signal.

Using the methodology described in this section, we have built a signal that only depends on F_0. In the next sections, we will present a methodology using such signals to conduct a perceptive evaluation using a F_0 transformation function described below.

3.4. Proposed subjective evaluation

The evaluation relies on ABX tests built from the sound samples obtained according to the previous method. Applying this method permits to take into account only the melody for each speaker. The subjective tests are realised by 10 testers and contain 40 test steps. For a whole test, at most 80 sentences for each speaker are necessary.

The principle of an ABX test is to present three sound samples A, B and X at each step of the test. The tester, once he has listened to these samples, has to answer the following question: is sample X may be paired with A or B ? There is no intermediate answer. The presentation order of samples A and B is random. Another important point is that the test does not include twice the same sentence, ie. with the same linguistic content. Then the non-parallel framework of the experiments is guaranteed.

Two experiments are conducted: the first one aims to evaluate the ability of listeners to discriminate between voices only thanks to their melody. To build this test, the preceding protocol is applied. Moreover, sample X can be derived from any of the two voices and the construction of the test insures that both voices has the same probability to occur. The second experiment evaluates the performance of a transformation function example using a ABX-type test where sample X stems from a fundamental frequency derived from a transformation function from A to B.

4. Experimental protocol

4.1. Data

The experiments are done on a couple of corpora in French that correspond to two male speakers. The first voice, called S_1, is constituted by records of short fantastic stories read in an expressive manner. The second speaker, S_2, is a corpus of read speech used in a text-to-speech system. Indeed, these two voices have a quite different elocution style.

Table 1: Confusion matrix for the discrimination test.The lines indicate the corpus from which sample X is extracted, the columns correspond to the answers of testers.

X \ Answer	Speaker S_1	Speaker S_2
Speaker S_1	80.0%	20.0%
Speaker S_2	12.0%	88.0%

For both corpora, a segmentation process into phones is carried out in an automatic manner [6]. The fundamental frequency, F_0, has been analysed automatically with the help of the YIN method [7]. Each phonetic sequence is segmented into syllables. F_0 contours are first interpolated and then smoothed before extracting the F_0 vectors.

For each voice, $10,000$ syllables are used for GMM training and $3,000$ syllables for validation.

4.2. Transformation function example

The transformation function used to illustrate the proposed evaluation methodology is taken from [3] and is aimed at adapting the prosody of speaker S_1 to the one of speaker S_2. It is based on an adaptation process using MLLR originally proposed by Gales & Woodland [8].

This method relies on a stylisation step in order to obtain constant sized F_0 vectors at the syllable level. Here, we only use the F_0 part of the prosodic vectors. The syllables are represented by three-dimensional vectors. Each coordinate of a F_0 vector respectively stands for the F_0 value at 10%, 50% et 90% of the syllable time support.

Next, a 32-component GMM $\mathcal{M}_{\mathbf{v}_1}$ is learnt using the source speaker data $\mathbf{s}_1$ and is then adapted using target speaker data $\mathbf{s}_2$. This adaptation results in an modified GMM $\widehat{\mathcal{M}_{\mathbf{s}_1}}$ which is used to build a transformation function providing the adapted data $\widehat{\mathbf{s}_1}$ as follows:

$$\widehat{\mathbf{s}_1} \; = \; \sum_{m=1}^{32} P(m|\mathbf{s}_1)(\widehat{\mathbf{W}}_m \xi_m) \qquad (3)$$

where ξ_m is the extended mean vector of the component m of $\mathcal{M}_{\mathbf{s}_1}$ and $\widehat{\mathbf{W}}_m$ is the adaptation matrix for this component.

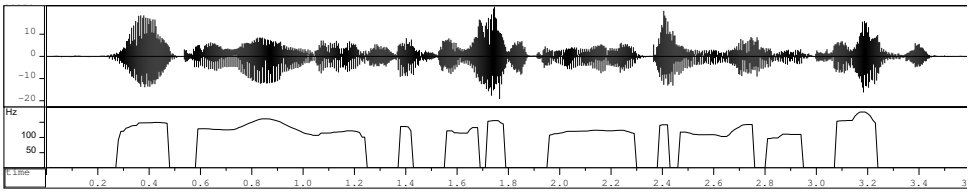
The regeneration of an F_0 curve from the sequence of tranformed vectors insures unvoiced segments are respected.

5. Results and discussion

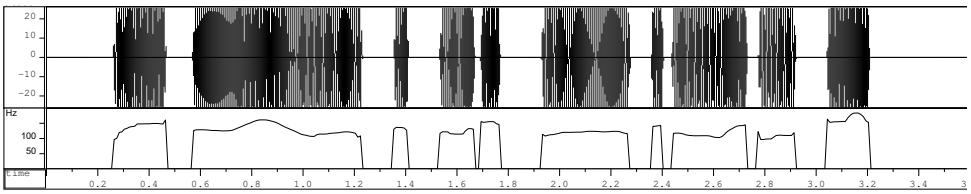
In this paragraph, the results of the two perceptive experiments are introduced. The first experiment is aimed at guaranteeing a significant difference exists between the melody of the two voices. The second one focuses on the evaluation of a transformation function example.

5.1. Preliminary discrimination test

The results of this experiment are summarized in Table 1 as a confusion matrix. We can notice that the difference between the voices is quite strong at a melodic level. In more than 80% of the cases, listeners succeed in discriminating the two voices. That means discriminating voices from only their melody can be achieved efficiently. Moreover, we can observe that samples stemming from the read speech corpus S_2 are less often wrongly paired with the expressive voice S_1 than the samples from S_1 with S_2.



(a) Original waveform and fundamental fraquency contour corresponding to the sentence “mais tout le monde dans l’ancien Far West ne menait pas une vie aventureuse !”.



(b) Waveform obtained using the chirp signal methodology and resultant F_0 contour. The F_0 contour drawn on this figure is obtained from the generated waveform.

Figure 1: Generation of a stimulus from a waveform. First F_0 analysis is done and then used to build a waveform that is independant from the segmental component.

Table 2: Results of the perceptive evaluation for MLLR based transformation method.

Answer / X	Source speaker S_1	Target speaker S_2
$\widehat{\mathbf{s}_1}$	58.7%	41.3%

Let us recall that this experiment is done in a completely non-parallel framework. In particular, that means testers cannot rely on a direct comparison between sentences melody and rhythm. The difficulty for a tester is then to find clues in melody that enable the identification of a speaker.

5.2. Evaluation of the transformation

Being done that the two voices can be discriminated by their melody, we now can evaluate a transformation function in order to estimate its position between source and target voices. Then the second step of the evaluation consists in replacing the X list by transformed samples $\widehat{\mathbf{s}_1}$. The transformation function that is evaluated using the proposed methodology is an MLLR based function. The size of the test is the same as in the previous one.

The result of this perceptive test is summarized in Table 2. We can notice that the result shows some proximity between the transformed and the source melodies. Nevertheless, the tested transformation induces the tester in error approximatively one time out of two. We can put this result into perspective as the proposed transformation function uses a poor F_0 stylisation method and implies the modification of the shape of the contours. The latter point can cause coherence errors for F_0 contours at the sentence level.

6. Conclusion

In the field of prosody transformation, evaluation is a crucial and difficult point. First, we have presented a technique based on chirp signals that gives the possibility of comparing voices only using the melody. The proposed method also insures the generated waveform has a continuous phase and guarantees the absence of audible artifacts. Secondly, we have introduced a non-parallel perceptive evaluation protocol that appears to be efficient to compare the melody of two voices. This protocol is also applied to an example of transformation function. The results show that although the transformed melody is closer to the source speaker, 40% of the tranformed data are attributed to the target speaker rather than the source one.

Finally, the proposed method can be also applied to parallel transformation functions so they can be compared to non-parallel functions in a common framework. Future works should focus on the assessment of the influence of test sentence duration and voiced/unvoiced transitions on this evaluation methodology.

7. References

[1] D. T. Chappell and J. H. L. Hansen, “Speaker-specific pitch contour modeling and modification,” in *Proc. IEEE ICASSP*, vol. 2, 1998, pp. 885–888.

[2] E. E. Helander and J. Nurminen, “A novel method for prosody prediction in voice conversion,” in *Proc. IEEE ICASSP*, vol. 4, 2007, pp. 509–512.

[3] D. Lolive, N. Barbot, and O. Boeffard, “Pitch and duration transformation with non-parallel data,” in *Proc. Speech Prosody*, 2008, pp. 111–114.

[4] T. Ceyssens, W. Verhelst, and P. Wambacq, “On the construction of a pitch conversion system,” in *Proc. EUSIPCO*, 2002, pp. 1301–1304.

[5] Z. Hanzlicek and J. Matousek, “F0 transformation within the voice conversion framework,” in *Proc. Interspeech*, 2007, pp. 1961–1964.

[6] L. Charonnat, G. Vidal, and O. Boeffard, “Automatic phone segmentation of expressive speech,” in *Proc. LREC*, 2008.

[7] A. de Cheveigne and H. Kawahara, “Yin, a fundamental frequency estimator for speech and music,” *Journal of the Acoustical Society of America*, vol. 111, pp. 1917–1930, 2002.

[8] M. J. Gales and P. C. Woodland, “Mean and variance adaptation within the mllr framework,” *Computer Speech & Language*, vol. 10, pp. 249–264, 1996.