Is the Fujisaki model a suitable (prosodic) model for the voice-conversion task?

Jan Schwarz, Martin Tran, Ulrich Heute

Institute for Circuit and System Theory (LNS), Faculty of Engineering,
Christian-Albrechts-University of Kiel, Germany, Email: {js, mt, uh}@tf.uni-kiel.de

Introduction

Voice conversion (VC) aims to transform the voice of one speaker (source) in such a way that the converted voice sounds as if it was uttered by another speaker (target). The meaning and content of the speech are not changed. Nowadays, many applications for the VC-task exist. An important application is a customised text-to-speech (TTS-) system which gives the ability to build *corporate identities* quickly and inexpensively by modifying the speech corpus of the TTS-system and thus the sound of the voice. Corporate identities are interesting for companies that like to represent the firm by one specific voice in public. They are a new form of advertisement.

VC can also be used to create special characters' voices for the movie industry or to "keep" the voice of an actor in different languages. The latter case aims to retain the speaker's identity in speech-to-speech translation-scenarios.

Different approaches with respect to voice conversion have been proposed in the last decades. From the technical point of view, statistical models are interesting as presented in [1] and [2]. The reason is in the characteristics of the speech signal. Speech varies from person to person strongly due to being related to the emotion of any person and, thus, expressing joy, sorrow, or anger. Furthermore, speech represents the mental attitude of a speaker by indicating whether he/she expresses ridicule or suprise. Therefore, using any rule-based approach as introduced in [3] and [4] seems to be inappropriate. Rules cannot model all nuances of speech and thus will lead to limitations.

The transformed voice can only sound natural, if all characteristics relevant for the true target speaker are included. Within VC-systems, a main problem is the mapping of the prosody which is one of the essential features. The prosody describes the rhythm and the intonation of speech so that it differs from speaker to speaker. In addition the prosody includes information on the stress as well as the lengthening and shortening of words and sentences.

In [5] it is shown that modelling the prosody will increase the quality of a VC-system with respect to the identity of the transformed voice and the true target speaker. A classification and regression tree is used to build a prosodic model which maps the prosody of source and target by using a codebook. This approach is rule-based and thus cannot be effective in general due to the characteristics of any speech signal.

Therefore, a statistical prosodic model has been developed for the VC-task [6]. The model uses Gaussian-Mixture Models (GMMs) trained for concatenated phonemes, so-called "multi-phones". It shows promising results in converting the fundamental frequency (F_0) as well as the duration. However, in some cases discontinuities occur in the fundamental-frequency (F_0-) contour which lead to significant distortions in the transformed voice. In such cases the corresponding GMM of any multi-phone was not trained sufficiently. In addition, the model does not have the ability to change the accentuation (stress) of any syllable or word.

This contribution introduces a parametric prosodic model for voice conversion. It is based on the Fujisaki model [7] which was developed to model the F_0-contour of Japanese utterances. In 1998, Mixdorff [8] modified the Fujisaki model to match the German language. The model was used within a TTS-system to generate a synthetic F_0-contour that includes different levels of accentuation. In this contribution the modified Fujisaki model of Mixdorff is analysed with respect to its suitability for the voice-conversion task.

The paper is organised as follows. First, prosodic models known from literature are compared concerning their use within VC-systems. Then, a modified Fujisaki model for the VC-task is introduced. It is used to convert the intonation of arbitrary sentences which are discussed afterwards. Finally, conclusions are drawn for this contribution.

Prosodic models

The term *prosodic model* is not used consistently in the literature due to intonation models that cover more than the intonation or the tonal (melodic) aspects. However, phoneticians distinguish between three levels of representing prosodic events, i.e. the perceptual, the linguistic and the acoustic level [9]. Perceptual models try to represent the prosody as heard by the listener. They include information about the perception, like the tone pitch. In contrast to perceptual models, the linguistic level models the prosody of an utterance as a sequence of abstract units, signs, or symbols, some of which have a communicative function in speech, while others may just fulfil syntactic requirements [9]. Thereby, the linguistic model is a structural interpretation of the data which results from the analysis of prosodic data by a linguist. Finally, acoustic models are derived from a parametrisation and analysis of the speech signal. The prosody is described by the fundamental frequency, duration, and/or the amplitude. The Fujisaki model which will be explained in the next section belongs to the class of acoustic models.

It has to be differentiated between the use of a prosodic model in VC as well as in phonetics. In VC, a prosodic model is a parametrisation of a speech signal by prosodic features that allow a conversion of the voice under several conditions [6]. Restrictions can either be given by the algorithm that is used to extract parameters to be modified or by the amount of data that is required to train a model, especially in the case of a statistical prosodic model. Furthermore, the question of how to map the extracted parameters from one speaker to another has to be answered.

The answer of the last question is important because it can differ from the original voice-conversion task. VC aims to transform the voice from a source speaker to sound as if it was spoken by the target speaker. In contrast to this, the direction of transforming the prosody is not fixed, i.e. the aim could either be to model the prosody of the target speaker or to keep the prosody of the source speaker. The first case reflects the direction of VC which aims to model all aspects of the target speaker and thus includes the prosody of the target speaker. But, the second case seems to be the better and appropriate way of transforming the prosody because it keeps the current speaking style and also the emotions of the source speaker. Keeping the source's prosody might be strange at first, but it is required to prevent situations in which the VC-system was trained with neutral (unemotional) utterances while the voice to be converted sounds sad or happy. In addition, not keeping the prosody of the source speaker can lead to ambiguity. Especially in German, some words (and even phrases) exist that do not differ in their writing but will have a different meaning depending on their accentuation. If the prosody of the target speaker would be kept as described in the first case, this could lead to wrong statements. Thus, the source speaker determines the prosody.

Fujisaki model in voice conversion

The Fujisaki model, originally developed for Japanese utterances [7], but also analysed for German [8], represents the prosody of an utterances by means of the F_0-contour. The F_0-contour can be described by two kinds of components [7]: The first, slowly varying components, named *phrase components*, which may or may not show slight initial rise and then decay towards an asymptotic baseline frequency F_b. The other components, named *accent components*, are local peaks or plateaus. The phrase components are the responses of a second-order linear system described by its impulse-response function $G_\mathrm{p}(t)$, whereas the phrase commands are given by a set of $i = \{1, 2, .., I\}$ impulses with different magnitudes A_{p_i}. The accent components are the responses of another second-order linear system with impulse-response function $G_\mathrm{a}(t)$, and the accent commands are a set of $j = \{1, 2, ..., J\}$ step functions with amplitudes A_{a_j}. If the Fujisaki-model parameters A_{a_j}, A_{p_i}, $G_\mathrm{a}(t)$ and $G_\mathrm{p}(t)$ are known, an F_0-contour can be expressed by [7, Eq. (1)-(3)]

$$\begin{aligned}\ln(F_0(t)) = \ln(F_\mathrm{b}) + \sum_{i=1}^{I} A_{\mathrm{p}_i} G_\mathrm{p}(t - T_{\mathrm{p}_i}) \\ + \sum_{j=1}^{J} A_{\mathrm{a}_j} \left\{ G_\mathrm{a}(t - T_{1_j}) - G_\mathrm{a}(t - T_{2_j}) \right\}, \quad (1)\end{aligned}$$

$$G_\mathrm{p}(t) = \begin{cases} \alpha^2 t e^{-\alpha t}, & t \geq 0, \\ 0, & t < 0, \end{cases} \quad (2)$$

$$G_\mathrm{a}(t) = \begin{cases} \min\left\{1 - (1 + \beta t) \cdot e^{-\beta t}, \gamma\right\}, & t \geq 0, \\ 0, & t < 0, \end{cases} \quad (3)$$

where T_{p_i}, T_{1_j} and T_{2_j} are the timing of the ith phrase command, the onset of the jth accent command, and the end of the jth accent command, respectively. The "natural angular frequency" α of the phrase-control mechanism and the "natural angular frequency" β of the accent-control mechanism are assumed to be constant at least within an utterance, while the relative ceiling level γ of the accent components is set equal to 0.9 [7]. The Fujisaki model gives the ability to generate an F_0-contour that is so close to the measured F_0-contour that both are perceptually indistinguishable in synthetic speech [7]. Thus, Mixdorff [8] uses the Fujisaki model to generate F_0-contours for a TTS-system by estimating the model parameters from the input-text given to the TTS-system.

With respect to voice conversion the Fujisaki model gives the possibility to parametrise the fundamental frequency in a certain way. Generally, all VC-systems use a parametrisation of certain parameters, including F_0, to transform the voice of the source speaker in such a way that it sounds like uttered by the target speaker. Thereby F_0 is usually given and also modified by means of its contour, which is generally speaking, a vector of arbitrary numbers. In contrast to this, the Fujisaki model forms another level of abstraction by generating an F_0-contour with the help of Eq. (1) to (3) and a reduced set of parameters.

Furthermore, the Fujisaki model has a physiological and physical background concerning the larynx and the vocal cords. The larynx is surrounded by three muscles which cause a change in the length and thus the tension of the vocal cords. Depending on the interaction of the muscles, F_0 is changed. However, F_0 can be divided into two parts; the first one is nearly constant and therefore related to the baseline frequency F_b, while the second one represents small changes in the vocal-cord length and thus is time-varying. The time-varying component is further divided into a translational and a rotational movement. Since the translational movement has a much larger time constant than the rotational movement, the former is used to indicate global phenomena such as phrasing, while the latter indicates local phenomena such as word accents [7].

Referring to VC the physiological interpretation of the Fujisaki model gives the ability to change the linguistic meaning of an utterance, by means of modifying the F_0-contour and the underlying parameters, respectively. Separating the F_0-contour into components allows the extraction of the accentuation, namely the accent components. Furthermore, the speaker-dependent phrasing is preserved which is given by F_b and the phrase components. Resynthesising the F_0-contour using F_b and the phrase components only, lead to a natural sounding, but "flat" voice. The importance of the Fujisaki model for the VC-task, and especially the prosody conversion, lies in the ability to change the relative prominence and thus the intonation of a syllable within a word or a group of words. The accentuation of any (target) speaker can be modified in such a way that it matches the accentuation and thus the prosody of the source speaker. In other words, the accent components of the target speaker are replaced by the accent components of the source speaker using a non-linear mapping. F_b and the target's phrase components are kept to resynthesize the F_0-contour.

Conversion of the intonation

In contrast to a TTS-system, a VC-system does not contain any textual information concerning the output voice. Thus, the Fujisaki-model parameters have to be extracted from the source's audio signal completely and cannot be verified by an underlying text (linguistic information). The extraction of the model parameters is based on [10] but slightly modified to match the VC-task.

The Fujisaki-model parameters are extracted from the F_0-contour solely. Therefore, Boersma's pitch-tracking algorithm [11] is used which is based on the autocorrelation method. Since the extracted F_0-contour contains macro- and microprosodic information, but the Fujisaki model explicitly deals with macroprosodic effects only, a smoothing has to be performed. In this contribution the modelling-melody (Momel) algorithm given by [12] is used to approximate the extracted F_0-contour by a quadratic spline interpolation. So, microprosodic effects which are uncontrollable variations of the F_0-contour are eliminated. In the following the smoothed F_0-contour will be denoted by $\overline{F}_0$-contour.

Known from the physiological background of the Fujisaki model, the (smoothed) $\overline{F}_0$-contour includes information related to the accent and phrase components. Since the accent components are highly fluctuating, whereas the phrase components vary slowly over time, the $\overline{F}_0$-contour is separated into a high-frequency component (HFC) and a low-frequency component (LFC) [10]. The accent components roughly correspond to the HFC while the phrase components are represented by the LFC. As in [10] a high-pass filter with cut-off frequency equal to 0.5Hz is used to extract the HFC from the $\overline{F}_0$-contour. The LFC derives from the HFC by subtracting the HFC from the $\overline{F}_0$-contour. Besides the sum of phrase components, the LFC contains the speaker-dependent baseline frequency F_b. F_b is set to the overall minimum of the LFC.

Since local maxima of the HFC correspond to the accent components and their amplitudes A_{a_j} roughly, succeeding (local) minima of the HFC are used to define time segments. The time segments are related to the onset time T_{1_j} and the end time T_{2_j} required for $G_\mathrm{a}(t)$ given by Eq. (3). It has to be noted that the $\overline{F}_0$-contour sustains in cases of a question or a non-terminal statement at a high level in the end of the utterance. The final accent would be missed due to a missing local minimum. Therefore, the last (detected) local minimum will be shifted to the end of the utterance, if the $\overline{F}_0$-contour does not decline to its lowest level.

Similar to determining accent components, phrase components are extracted from the LFC by performing a local maximum and local minimum search. The local maxima correspond to the magnitudes A_{p_i} of the phrase components whereas the local minima define the time segments of phrase components given by T_{p_i} and Eq. (2). The first phrase component is assumed to start at the beginning of the utterance, and the last phrase component is assumed to end at the end of the utterance even though a local minimum is missing at the beginning or end, respectively. According to [13] a phrase component will at least have a duration of 750ms. A further classification into shorter prosodic phrase components is not motivated linguistically.

Up to this point, Fujisaki-model parameters are extracted that could roughly approximate the given $\overline{F}_0$-contour. Thus, the parameters are adjusted recursively in the least-squares sense to match the $\overline{F}_0$-contour. The iteration stops if the overall error between the parameterised (Fujisaki) $\widehat{F}_0$-contour and the $\overline{F}_0$-contour is smaller than 5%. However, adjusting the model parameters can lead to overfitting so that the model parameters are not interpretable linguistically. To overcome this problem, fitting the model parameters is done segmentwise, i.e., the number of phrase components are given by the number of extracted time segments and will not be changed. On the contrary, accent components are allowed to be cancelled out if they are meaningless, i.e. having a duration smaller than 50ms [8].

If the Fujisaki-model parameters are extracted for the source and target speaker respectively, the intonation could be converted from the source to the target speaker. In contrast to modifying the complete F_0-contour, the Fujisaki model allows to transform single parameters by performing a mapping. The mapping has to be done in a non-linear way due to differing speaking styles and speaking rates of source and target speaker. It is carried out using a dynamic time-warping approach. During the conversion, the phrase components of the target speaker will be kept whereas the accent components will be substituted by the source's ones.

Results

The German utterance "Ein Junggeselle ist ein Mann, dem zum Glück noch die Frau fehlt." has two different meanings depending on the accentuation of the word

"Glück" (engl. luck). Without emphasising "Glück" the sentence can be translated into "A bachelor is a man who is missing a girlfriend to be happy.". Fig. 1(a) shows the corresponding F_0-contour (gray), the $\widehat{F}_0$-contour given by the Fujisaki model (dashed) and the phrase components (dash-dotted). The F_0-contour declines slightly to a baseline frequency and does not show any noticeable peaks. The corresponding accent and phrase commands given in Fig. 1(b) certify this behaviour by staying at a low level. In contrast, if the word "Glück" is accentuated the meaning of the utterance is changed into "A bachelor is a man who is lucky not to be bonded to a girlfriend.". The F_0-contour (Fig. 1(e)) shows a perceivable peak at 1.6s indicating an accentuation. Additionally, the corresponding accent command at 1.5s given in Fig. 1(d) illustrates the accentuation of "Glück".

Mapping the source's intonation into the target ones lead to an accent-command structure as given in Fig. 1(c). The word "Glück" is not emphasised anymore. The belonging accent commands stay at a low level and correspond to the commands given by the source (cf. Fig. 1(b)). They are shifted in time due to a non-linear mapping.

Conclusion and future work

This contribution focuses on the Fujisaki model and its use within the VC-task. The model can be used to modify the accentuation of any target speaker, if the model parameters of the source are mapped non-linear. Thus, the prosody of the target can be changed.

However, the Fujisaki model was not tested in a VC-system so far. In a next step it will be tested in such a scenario.

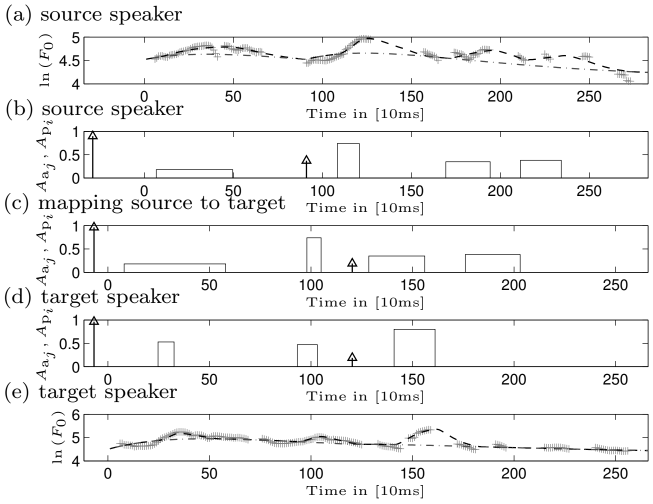

Figure 1: a./e.) F_0-contour (gray), phrase components (dash-dotted) and Fujisaki $\widehat{F}_0$-contour (dashed) of source and target respectively, b.-d.) accent commands (rectangular) and phrase components (arrows) of source, transformed and target speaker, respectively

References

[1] Stylianou, Y; Cappé, O.; Moulines, E.: Continuous probabilistic transform for voice conversion. IEEE Transactions on Speech and Audio Processing (1998), Vol. 6, No. 2, 131-142

[2] Kain, A.; Macon, M.: Spectral voice conversion for text-to-speech synthesis. IEEE International Conference on Acoustics, Speech and Signal Processing (1998), Seattle, USA, 285-288

[3] Abe, M.; Nakamura, S.; Shikano, K.; Kuwabara, H.: Voice conversion through vector quantization. Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (1988), New York, USA, 655-658

[4] Kuwabara, H.; Sagisaka, Y.: Acoustic characteristics of speaker individuality: Control and conversion. Speech Communication (1995), Vol. 16, No. 2, 165-173

[5] Helander, E. E.; Nurminen, J.: A Novel Method For Prosody Prediction in Voice Conversion. Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (2007), Vol. 4, 509-512

[6] Schwarz, J.; Heute, U.: A Statistical Prosodic Model for Voice Conversion. Proc. of Acoustics '08 Paris, 155. Meeting of the Acoustical Society of America (2008), Paris, France, 1267-1272

[7] Fujisaki, H.: Information, Prosody, and Modeling – with Emphasis on Tonal Features of Speech. Proc. of Speech Prosody (2004), Keynote speech, Nara, Japan, 1-10

[8] Mixdorff, H.: Intonation Patterns of German – Model-based Quantitative Analysis and Synthesis of F_0 contours. Fakultät für Elektrotechnik, Technische Universität Dresden, Germany, PhD Thesis, 1998

[9] Dutoit, T.: An Introduction to Text-to-Speech Synthesis. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997

[10] Mixdorff, H.: A novel approach to the fully automatic extraction of Fujisaki model parameters. Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (2000), Istanbul, Turkey, Vol. 3, 1281-1284

[11] Boersma, P.: Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. IFA Proceedings (1993), Institute of Phonetic Sciences, University of Amsterdam, The Netherlands, Vol. 17, 97-110

[12] Hirst, D.; Espesser, R.: Automatic Modelling of Fundamental Frequency using a quadratic spline function. Travaux de l'Institut de phonétique d'Aix (1993), Vol. 15, 75-85

[13] Mixdorff, H.: An Integrated Approach to Modeling German Prosody. Fakultät für Elektrotechnik, Technische Universität Dresden, Germany, 2002