euronoise

A Statistical Prosodic Model for Voice Conversion

Jan Schwarz and Ulrich Heute

Institute for Circuit and System Theory, Christian-Albrechts-University of Kiel, Kaiserstrasse 2, 24143 Kiel, Germany
js@tf.uni-kiel.de

In this contribution a statistical prosodic model for the voice-conversion task is presented. Voice conversion (VC) aims to transform the voice of one speaker in such a way that the converted voice sounds as if it was uttered by another speaker. The meaning and content of speech are not changed. Nowadays, VC-systems suffer from a poor naturalness and quality of the transformed voice due to not including any prosodic model. Therefore, a statistical prosodic model is introduced that is based on Gaussian-Mixture Models (GMMs). The GMMs are trained for the fundamental frequency f_0 and the duration T of diphones. To ensure sufficient training data for the GMMs, the diphones are grouped into classes related to the International Phonetic Alphabet.

1 Introduction

Nowadays Text-to-Speech (TTS-) systems try to increase their acceptance by adapting the system to the user and his/her usage. Different approaches exist, but they are directed to the sound quality of the synthesized voice mainly. In some approaches the global aim is to build synthesised voices that sound "natural" and "human-like", while other approaches try to create voices that sound like a specific person. The latter case is known as so-called "custom voices" [1] or "corporate identities" [2]. Corporate identities are interesting for companies especially due to giving the ability to represent the firm by one specific voice in public. Whenever a person hears that voice it will be reflected to the company and its products. So, corporate identities are a new form of advertisement.

Naturalness and also intelligibility can be reached by using recordings of human voices. Short speech components are concatenated to form the output voice. Using such a technique is time consuming and also costly. In addition, the vocabulary is limited due to giving the opportunity of concatenation for known words or syllables only. If a word or a syllable is missing, it will be left out or created artificially by using specified rules. Thus, the generated voice sounds unnatural and robotic due to using a different flow of words that stands in contrast to the speech flow of the original speaker.

A technique that has the potential to create a personalised TTS system and saves ressources, is voice conversion (VC). In VC the aim is to transform the voice from one speaker (source) to sound as if it was spoken by another speaker (target) without changing the meaning or the content of speech. Usually, a set of training material is recorded from the source and target speakers, and one or more conversion models are trained. In the literature different approaches in respect to voice conversion have been proposed. From the technical point of view, statistical models are interesting as presented in [3] and [4]. The reason for this is in the characteristics of the speech signal. Speech varies from person to person strongly because it is related to the emotion of any person and, thus, expresses 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 [5] and [6] seem not to make sense: Rules cannot model all nuances of speech and thus will lead to limitations.

The transformed voice can only sound natural, if it includes all characteristics relevant for the true target speaker. 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 [7], Helander et.al. show that modelling the prosody will increase the quality of a VC-system in respect to the identity of the transformed voice and the true target speaker. In her work, Helander uses a classification and regression tree to build a prosodic model which maps the prosody of source and target by using a codebook. That approach is interesting as a first approximation, but a rule-based model, as given by a codebook mapping, cannot be effective in general due to the characteristics of any speech signal. Speech cannot be modeled by rules so that an enhancement should result from a statistical prosodic model.

This contribution introduces a statistical prosodic model for voice conversion. The approach is based on Gaussian-Mixture Models (GMMs) that are trained for the fundamental frequency f_0 and the duration T of diphones, respectively. To ensure sufficient data for the training of the GMMs, the diphones are separated into classes that are related to the International Phonetic Alphabet (IPA) [8].

The paper is organised as follows. In Section 2 prosodic models known from the literature are compared in respect to their suitability for voice conversion. Then a statistical model for voice conversion is presented. Section 3 focuses on experiments which were carried out to determine the optimal parameters for the presented model. The experimental results are discussed in Section 4. Section 5 concludes this contribution.

2 Prosodic models

2.1 Prosodic models in phonetics

Within the literature, the term "prosodic model" is not used consistently. Some "intonation models" exist that describe more than the intonation or the tonal (melodic) aspects and thus can be said to be prosodic models. However, as known from phonetics, prosodic events can be studied at three levels of representations, i.e., the acoustic level, the perceptual level, and the linguistic level. Acoustic models are derived from a parametrisation and analysis of the speech signal. The prosody is described by the pitch f_0, duration T and/or the amplitude. The so-called Fujisaki model [9] belongs to the class of acoustical models and gives a representation of the prosody in terms of the f_0-contour. In contrast to the acoustic model, perceptual models represent prosodic events as heard by the listener. They give infor-

mation about the perception, like the tone pitch. Such a model is given by [10]. Finally, the linguistic level represents the prosody of an utterance as a sequence of abstract units (signs, symbols), some of which have a communicative function in speech, while others may just fulfil syntactic requirements [11]. Thereby, the linguistic model is a structural interpretation of the data, which results from the analysis of prosodic data by a linguist. The so-called "Kieler Intonation Model (KIM)" [12] belongs to the class of linguistic models.

2.2 Prosodic model in voice conversion

In respect to voice conversion, the term prosodic model has to be distinguished from its use within 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. In this context restrictions are given by the algorithm that is used to map the prosody of the source to that of the target speaker, which should be a statistical approach in this case, and the amount of data that is required to train the GMMs. Furthermore, the question of how to cope with many-to-one and one-to-many mappings has to be answered. To describe the prosody in terms of features, f_0 and T are used first of all.

The presented statistical model is based on M Gaussian-Mixture Models that are used to describe an arbitrary conditional probability function $P(C_i|\,\boldsymbol{x})$. The conditional probability $P(C_i|\,\boldsymbol{x})$ describes that an observation vector $\boldsymbol{x}$ belongs to the acoustical classes C_i of the GMM. A class $C_i,\ i \in \{1, 2, ..., M\}$, is given by a $\mathcal{D}$-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ according to

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{\left(\sqrt{2\pi}\right)^{\mathcal{D}}} |\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_i)\right\}. \quad (1)$$

In Eq. (1) the exponent $^{\mathrm{T}}$ indicates a transposition. The mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ can be estimated using the expectation-maximisation (EM) algorithm that is described in [13, 14] in more detail.

With the help of Bayes' rule, the conditional probability $P(C_i|\,\boldsymbol{x})$ and thus the GMM can be given by

$$P(C_i|\,\boldsymbol{x}) = \frac{\alpha_i \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i))}{\sum_{j=1}^{M} \alpha_j \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}, \quad (2)$$

where $\alpha_i,\ i \in \{1, ..., M\}$, are the mixture weights representing the statistical frequency of each of the M classes in the observation. They have to fulfil the condition

$$\sum_{i=1}^{M} \alpha_i = 1. \quad (3)$$

Using Eq. (2) leads to the conversion function $G(\boldsymbol{x})$ which can be given in analogy to [3] as

$$G(\boldsymbol{x}) = \sum_{i=1}^{M} P(C_i|\,\boldsymbol{x})\,[\boldsymbol{\nu}_i + \max\{\Phi_i, \epsilon\}(\boldsymbol{x}-\boldsymbol{\mu}_i)], \quad (4)$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\nu}_i$ are the mean vectors of the source and target speakers, respectively. Φ_i is a correction term which is derived from the variance $\sigma^2_{\boldsymbol{x}}$ of the source-feature vector $\boldsymbol{x}$ and the variance $\sigma^2_{\boldsymbol{y}}$ of the target-feature vector $\boldsymbol{y}$ according to

$$\Phi_i = \frac{\mathrm{E}\left\{(\boldsymbol{x}-\boldsymbol{\mu}_i)^2\right\}}{\mathrm{E}\left\{(\boldsymbol{y}-\boldsymbol{\nu}_i)^2\right\}} = \frac{\sigma^2_{\boldsymbol{x}}}{\sigma^2_{\boldsymbol{y}}}, \quad (5)$$

where $\mathrm{E}\{\cdot\}$ indicates the expectation operator. Φ_i is limited by the factor $\epsilon,\ \epsilon > 0$, to prevent a too high influence on the term $(\boldsymbol{x}-\boldsymbol{\mu}_i)$.

2.3 Conditions to be met

In the presented method, GMMs are used to build a conversion function $G(\boldsymbol{x})$. GMMs are commonly trained "from scratch" using a relatively large amount of aligned training data. The training data can be either "parallel", this means both the source and target speakers read the same text, or "unparallel". In this contribution, it is assumed that the training data is parallel.

Although a parallel set of training data is used, the utterances of source and target speakers are not aligned perfectly. The speaking rates of both speakers differ normally, so that corresponding data-frames cannot be mapped directly. An alignment has to be performed. The simplest alignment technique is linear interpolation that assumes a global variation in the speaking rate. However, the speaking rate varies not only globally due to the speaking style and the emotions of a speaker. Thus, local variations have to be taken into account that lead to a non-linear interpolation. Non-linear warping can be obtained using the dynamic time warping (DTW) algorithm that finds an optimal path through a difference matrix computed between the source and target features.

The purpose of DTW is to generate a non-linear warping function of feature sequences along the time axis. This means a target frame can be mapped to more than one source frame. Also, a source frame can have more than one target frame mapped into it. This results in one-to-many and many-to-one mappings. If the speaking rate of source and target speakers differ significantly, this can lead to ambiguous training data. In that case the GMMs would be trained inaccurately.

Furthermore, it is essential that a high amount of training data is available to train the GMMs sufficiently. In this approach, it is assumed that multi-phones ($\mathrm{MPhones_i}$) represent the prosodic properties to be described by GMMs. In this context a $\mathrm{MPhone_i}$ is defined as a concatenation of i phones, where i = 1 is a phone ($\mathrm{MPhone_1}$), i = 2 a diphone ($\mathrm{MPhone_2}$) and so on. Generally, a single phone cannot represent any prosodic property due to its missing context. Thus, GMMs are trained for $\mathrm{MPhones_i}$, where $\mathrm{i} \geq 2$, to model the prosody. To overcome the problem of having too few training data to train the GMMs for one $\mathrm{MPhone_i}$, the data is grouped into seven classes related to the IPA. The classes are specified as follows: plosive, nasal, fricative, lateral, vowel, diphthong, and miscellaneous. Any combination of different classes is possible, if the index i of a $\mathrm{MPhone_i}$ is greater than 1.

In contrast to VC that transforms the voice from a source speaker to sound as if it was spoken by another speaker, the transformation of the prosody has to be

performed the other way around: The output voice shall sound like the true target speaker but the speaking style and also the emotions of the source speaker shall be kept. Thus, the source speaker determines the prosody. This is opposed to the voice-conversion task but required to prevent that the transformed voice sounds for example happy due to the given training data while the source speaker speaks sadly.

3 Experiments

The database consists of a set of the *Berlin sentences* taken from the German speech database "The Kiel Corpus of Read Speech" (KCoRS) [15] sampled at 16kHz. All the sentences have been manually labelled by experts so that these labels are assumed to be precise and taken for the determination of the duration T of the MPhones_i. The GMMs for f_0 and T are trained on 80 of 100 sentences for the speakers k04, k05, k06 and k65 (2 female, 2 male) using a different number M, $M \in \{1, 2, ..., 10\}$, of normal distributions to model the prosody. In addition, it is analysed, if MPhones_i to the base i equal to two and three are appropriate to model the prosody. Furthermore, the correction term Φ_i and its influence on the conversion of the prosody is analysed by limiting the variance of f_0 and T.

During the experiments the fundamental frequency f_0 is determined using the proposed algorithm by Boersma [16] which is based on the autocorrelation method. Pauses and/or silent frames are not included in the training of the GMMs. In addition, MPhones_i are not created over word boundaries. This means, the MPhones_i are provided from single words only, not including any pauses.

During the evaluation of the presented model, 20 sentences from the Berlin sentences of the KCoRS, that were not included in the training data, are used to convert f_0 and T by using Eq. (4). The performance of the conversion was analysed by an informal subjective listening test and by f_0- and T-contours.

4 Results

Starting from the experiments described in Section 3, the figures 1 and 2 show results converting the prosody of identical genders (male-to-male, Fig. 1) and different genders (male-to-female, Fig. 2) using MPhones_2 and $M = 5$ GMMs. A conversion from female-to-male and from female-to-female is left out due to having the same aspects and problems as the presented conversions suffer from.

All figures show the f_0-contours of the source speaker (top), the target speaker (bottom) and the transformed f_0-contour (middle). In the case of a same-gender conversion (Fig. 1) the f_0-contours do not differ largely from source and target speakers, except in their time-alignment. Thus, a prosody conversion succeeds in most instances. However, if the GMMs are not well-trained for any MPhone_i combination due to being rare or even missing within the training data, a prosody-conversion fails. Discontinuities in the f_0-contour result which lead

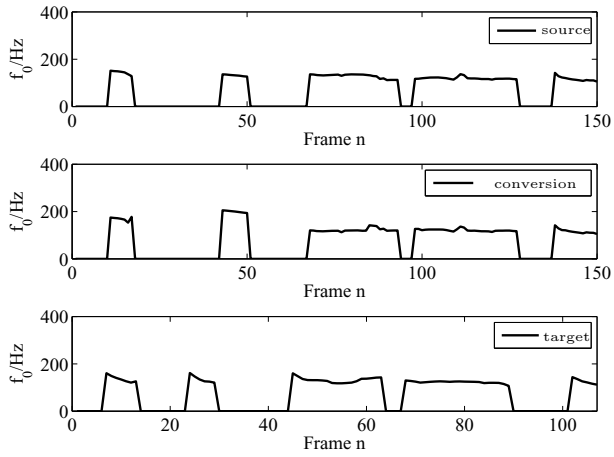

Figure 1: f_0-contours of source (male, k05) (top) and target speakers (male, k65) (bottom) as well as the transformed f_0-contour (middle), using $M = 5$ GMMs and MPhones_2

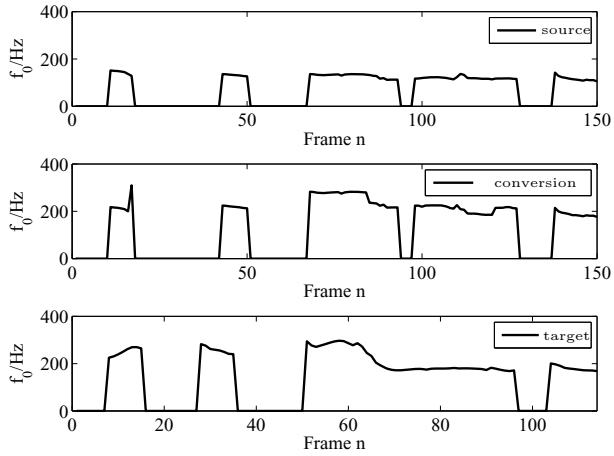

Figure 2: f_0-contours of source (k05, male) (top) and target speakers (k06, female) (bottom) as well as the transformed f_0-contour (middle), using $M = 5$ GMMs and MPhones_2

to a unnatural sound of the transformed voice. If MPhones_3 are used, the amount of data is smaller compared to MPhones_2 so that discontinuities occur more often. Thus, it is recommended to use MPhones_2.

If a cross-gender conversion of the prosody is performed (male-to-female, Fig. 2), the conversion does not succeed as well as for the same genders. This is due to the difference of the f_0-contours so that the characteristics of the target speaker will still have influence on the transformation. This can be seen from the converted f_0-contour in Fig. 2, where the characteristics of the converted f_0-contour go down starting from frame 80 while the source speaker does not have that strong descent.

In addition, the number M of GMMs has an influence on the conversion. In fact, the more GMMs are used to model any probability function, the better the approximation of the function will become. However, if a prosody conversion is performed this statement does not hold due to the following fact: The amount of training data is limited for many reasons. So, the approximation of a probability function as given by the GMMs

cannot be as precise as it should be. An error will occur that leads to oversmoothing of the GMMs. This will result in discontinuities in the converted f_0-contour which in turn influence the naturalness of the prosody. Figure 3 gives an example for a male-to-male conversion using $M = 8$ GMMs to model the prosody. The effect of oversmoothing can be seen clearly within the frames 75 to 125 of the converted f_0-contour due to the discontinuities in the f_0-characteristics.

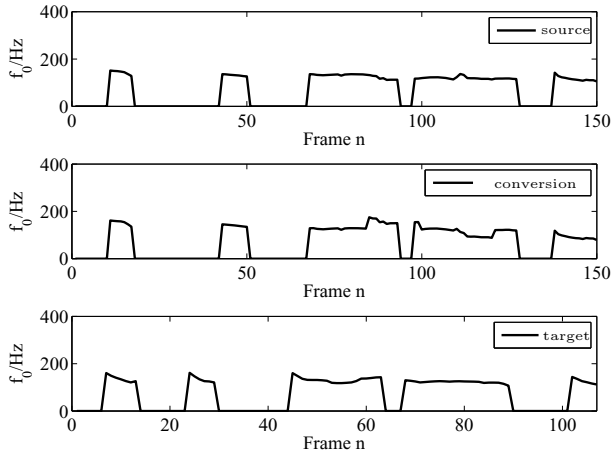

Figure 3: f_0-contours of source (k05, male) (top) and target speakers (k65, male) (bottom) as well as the transformed f_0-contour (middle), using $M = 8$ GMMs and MPhones_2

Furthermore, if the amount of training data is small but the number M of GMMs is large, the variance of each GMM will decrease. In the worst case, each sample of the training data will belong to one single GMM, and the variance will be zero. Thus, the GMM will become a dirac impulse and the conversion will be a single mapping of two data points. A codebook mapping will result which stands in contrast to the statement that a rule-based approach cannot include all aspects of a speech signal due to the variability of the voice. To overcome this problem, the factor ϵ is introduced in Eq. (4) which provides a minimal variance and thus guarantees the variability of the speech signal.

An analogous effect happens, if the number M of GMMs is too small. In this case the variance of the GMM will not become zero but it will be large. So, the variability within the speech signal will get lost due to creating one single GMM for different aspects of the prosody. This will also lead to discontinuities within the f_0-contour and will effect the naturalness of the converted voice.

During this study, a number of $M = 5$ GMMs performed best.

The speaking rates of two speakers differ normally due to their speaking style. The speaking rate varies locally so that GMMs for the duration T of MPhones_i are trained. The conversion of T turned out to be difficult in respect to the quality of the transformed voice. The used Synchronised Overlap-Add (SOLA) algorithm [17] can lead to discontinuities which are audible by distortions in the converted speech signal.

5 Conclusion

Generally, the prosody can be modeled in a statistical way by using GMMs. Such a model is suitable for voice conversion, if the model parameters T and f_0 are used. However, a statistical model depends on the amount of training data used to train the GMMs. Normally, the amount of data is strongly limited in voice conversion, so that strategies have to be invented to increase that amount. The presented model is based on MPhones_i, a concatenation of i phones, whose number of occurrence is limitted due to the training data. Thus, the GMMs would not be trained sufficiently, if GMMs would have been build for any single MPhone_i. So, the MPhones_i were grouped into classes in respect to the IPA and the GMMs trained on those groups.

The experiments show that GMMs for MPhones_2 can be generated sufficiently well and that the context, one of the fundamental requirements to model the prosody, is included. From the theoretical point of view, also GMMs for MPhones_3 should guarantee a prosody-conversion but the amount of the used training data is too small to train the GMMs sufficiently. In such cases, discontinuities in the converted f_0-contour occur that lead to an unnatural sound. Thus, MPhones_2 are recommended.

Furthermore, the experiments have shown that the number M of GMMs also influences the conversion-quality. If M is chosen too small or too large, discontinuities in the converted f_0-contour result which will decrease the sound-quality of the converted voice concerning naturalness. During the experiments, $M = 5$ performed best.

Finally, the used SOLA algorithm is one of the main drawbacks of the presented model. It is required to allow local changes in the speech rate depending on the MPhones_i. However, the algorithm leads to discontinuities which are audible by distortions in the converted speech signal.

In future work, the SOLA algorithm has either to be replaced by a different approach that will not effect the converted voice by any distortions, or some kind of smoothing has to be performed to avoid distortions.

Furthermore, the presented statistical prosodic model has to be expanded by a module that includes the stress. Indeed, the stress could be represented by a change of the amplitudes, but the perception of loudness is not exclusively related to the amplitude (and energy) at one point of the speech signal. It is also dependent on the duration of a speech fragment, in this case e.g. a MPhone_i. Moreover, the loudness is not noticed at one point, but relative to the loudness of other parts in the signal.

In addition, in future work the use of labels taken from the KCoRS has to be replaced by a method that can detect the corresponding classes from the IPA itself. A speech recogniser is able to detect and group single MPhones_i, but its complexity seems not to be suitable in the voice-conversion task. Moreover, a speech recogniser also requires a lot of data during the training so that this is the main reason why it is not reasonable for VC.

References

[1] R. Schmald, "Custom Voices" (in German), *Funkschau 15/2006*, 22-24 (2006)

[2] A. Kauffeisen, "Das sprechende Image" (in German), *Funkschau 23/2005*, 14-15 (2005)

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

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

[5] M. Abe, S. Nakamura, K. Shikano, H. Kuwabara, "Voice conversion through vector quantization", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New York, USA, 655-658 (1988).

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

[7] E.E. Helander, J. Nurminen, "A Novel Method For Prosody Prediction in Voice Conversion", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007), Vol. 4*, 509-512 (2007)

[8] International Phonetic Association, "Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet", *Cambridge University Press*, Cambridge, Great Britain (1999)

[9] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing", *In:* P.F. MacNeilage,"The production of speech", *Springer*, New York, 39-55 (1983)

[10] C. d'Alessandro, P. Mertens, "Automatic Pitch Contour Stylization Using a Model of Tonal Perception", *Computer, Speech, and Language, Vol. 9, Iss. 3*, 257-288 (1995)

[11] T. Dutoit, "An Introduction to Text-to-Speech Synthesis", *Kluwer Academic Publishers*, Dordrecht, The Netherlands (1997)

[12] K.J. Kohler, "Parametric Control of Prosodic Variables by Symbolic Input in TTS Synthesis", *In:* J.P.H. van Santen, R.W. Sproat, J.O. Olive, J. Hirschberg,"Progress in Speech Synthesis", *Springer*, New York, 459-476 (1997)

[13] L. Rabiner, B.-H. Juang, "Fundamentals of Speech Recognition", *Prentice Hall Signal Processing Series, Prentice-Hall Inc.*, New Jersey, USA (1993)

[14] A.P. Dempster, N.M. Laird, D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society, Series B (Methodological), Vol. 39, No. 1*, 1-38 (1977)

[15] K.J. Kohler, "The Kiel Corpus of Read Speech", *Institute of Phonetics and digital speech processing at the Christian-Albrechts University of Kiel*, Germany (1994)

[16] P. Boersma, "Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound", *Institute of Phonetic Sciences, University of Amsterdam, Proceedings 17*, The Netherlands, 97-110 (1993)

[17] D. Hejna, B.R. Musicus, "The SOLAFS Time-Scale Modification Algorithm", *BBN Technical Report*, University of Cambridge, Great Britain, (1991)