Article can be accessed online at http://www.publishingindia.com

Modified PSOLA-Genetic Algorithmbased Approach for Voice Re-Construction

Partha Sarthy Banerjee*, Uttam Kumar Roy**

Abstract

The process by which we try to reconstruct or regenerate a voice sample from a source sample or try to modify a source voice to a desirable voice, is called synthetic voice generation or artificial voice or voice conversion. The basic and conventional remedies to overcome this issue are based on training and applying conversion functions which generally require a suitable amount of pre-stored training data from both the source and the target speaker. The paper deals with a very crucial issue of achieving the required prosody, timber and some other unique voice templates by considerably reducing the dependence on the sample training dataset of voice. We needed to find out a way by which we can have templates of the "to be achieved voice" which are nearly same parametrically. This is achieved by assigning a marker to the target voice sample for training. A proper estimation of the transformation function can be made possible only by the above mentioned data. We can get the process done by pre-existing methods. In a nutshell, what we proposed is a system by which even in the scarce availability of training dataset, we can reach to a considerable amount of closeness of the target voice. Even though there is a disadvantage that to have higher precision and closer resemblance, we need to have clear idea of the system of spelling that a language uses.

Keyword: Artificial Voice, Prosody, Timber, Source Voice, Target Voice, Formant Structure

Introduction

While considering the synonymous words like voice, sound, speech, noise or music, all correspond to human's eternal instinct to perceive something by "LISTENING". Out of all the varieties of mechanical wave form widely available in the nature, the most intriguing is human voice. Un-denying the fact that soundwave forms as compared to other sources of perception are equally important, but it's the complexity and the dynamically varying nature of the human voice that fascinate us the most. The moment we start talking about voice or speech the developments that are associated with this arena which dawns in our thoughts, are speech synthesis and recognition, voice commands and interactive voice response and many more.

In the light of artificial intelligence where much of the effort that has been invested is in creating intelligence part for the robots for its articulation in movements, prediction abilities, computer vision, pattern recognition or even artificial thoughts, very less has been endeavoured for artificial voice. Now these are the aspects to name a few of them. For the time being it's the sheer human ingenuity and the basic transcendental quest to achieve the other utilities of speech or voice synthesis.

Review Work

Voice Reconstruction Process

The voice reconstruction process may be divided into a sequence of steps:

* Assistant Professor, Department of Computer Science & Engineering, Jaypee University of Engineering & Technology, Guna, Madhya Pradesh, India. E-mail: partha1010@gmail.com

** Assistant Professor, Department of Information Technology, Jadavpur University, Kolkata, West Bengal, India. E-mail: u_roy@itjusl.ac.in

Step 1: The initial step may be regarded as the pre-requisite step where the most unique features of both the source as well as the target dataset are traced. This phase is called the Analysis step. The values achieved are basically the parameters associated with the speaker identity like pitch, prosody and formant frequencies and a few more.

Step 2: In the second step we try to map the features of the source voice computed in the previous step to that of the "to be achieved voice" to as close a proximity as possible. This phase is controlled by a conversion rule obtained by a training phase.

Step 3: Synthesis: Last but not the least is the final phase where the modified parameters are used to synthesis or reconstruct the new speech which generally does have the target voice as well as the required prosody too if the module assists.

The researches on the crux topics of voice and speech are basically revolving around the paradoxical axis of speech or voice synthesis with application areas in the form of text to voice and vice-versa with stress on characterization and bifurcation of two or more voice samples. Some of the broad application areas are as enumerated.

Review of Speech Synthesis Technology

Unnatural or synthetic or artificial speech has been developed steadily over the last decades. Especially, the intelligibility has attained an adequate level for most of the user defined applications, viz. for communication impaired people. The intelligibility of artificial speech may also been hanced considerably with visual information. Speech synthesis may be classified as restricted in the form of messaging and at times unrestricted when it is text-to-speech processing. The first one is suitable for announcing and information systems while the latter is needed for example in applications for the visually impaired (Lemmetty and Karjalainen, 1999). In the sections to follow we are going to exemplify some of the topics.

Text-to-Phonetic Conversion

The basic problematic area encountered by any TTS system is the conversion of input text into linguistic representation, which is generally known as text-to-phonetic or grapheme-to-phoneme conversion. Another aspect to be noted is that every single language has got its own unique way of conversion and this leads to tremendous increment in the complexity level. In some languages, such as Finnish, the conversion is easier as written text almost corresponds to its pronunciation, whereas for English and most of the other languages the conversion is much more complicated. A correct pronunciation and prosodic flow is the outcome of strict coherence to a colossal amount of language axioms and rules. The correctness does not only depend on these facts and rules rather a huge set of exceptions too. The process basically begins with the preprocessing of the text to be converted and then an in-depth analysis of the data for a unique and correct pronunciation. The last step involves the proper computation of the prosodic features. A few of the important steps have been discussed here.

Text Preprocessing

The first step i. e. text preprocessing can be understood about its difficulty level from the following example where any English numeral, let's say121, may be at first read as one hundred and twenty one and 2014 as twenty fourteen if inferred as year or two thousand and fourteen for quantizing something for measurement. Some of the similar cases are the distinction between the any numeral and then stating pilot or people. The final area is the fractions and dates which are equally troublesome. 4/14 can be expanded as four-fourteenths (iffraction) or April Fourteenth (if date). The above mentioned problems are amongst a few of the problems encountered and the corresponding solution.

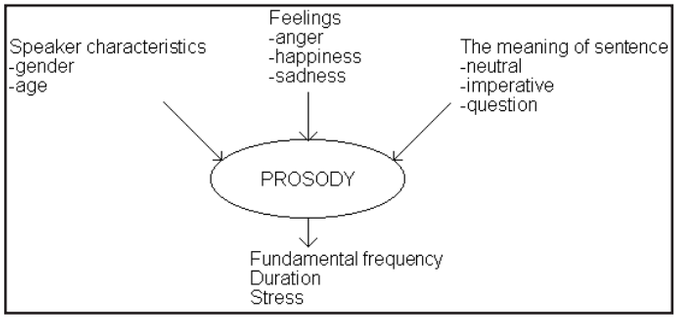
Pronunciation

The next important task is to find correct pronunciation for different relevant areas in the text. There are certain types of words that bear same spelling but have different meaning and sometimes different pronunciations also which are called homographs. Now such words are a big obstacle for the overall module. While considering an English word 'lead' we generally can end up with different pronunciations depending on its utility as a verb or noun, and between two noun senses. With all these kinds of words, some semantically helpful as well as informative data are highly essential to achieve correct pronunciation.

Prosody

The rhetorical flow of any word or its segment will definitely consider the correct intonation, proper stress at the punctuations and duration from written text is probably the most challenging problem for years to come. All these features when cumulatively considered are called prosodic features and may be considered as a melody stream, rhythmic flow and stress of the speech at the perceptual state. When the fundamental frequency or also at times which is the same the varying patterns in the pitch varies during the whole voiced segment of speech then it may be regarded as a particular intonation. The basic meaning of the uttered phrase and the emotional state of the speaker are some of the deciding factors for the prosodic characteristics. The dependencies of prosodic variations are shown in Figure 1. The ironic situation is that any information in written or textual format doesn't carry the traits of these qualities (Lemmetty and Karjalainen, 1999).

Figure 1: Prosodic Dependencies



The forthcoming segments will deal with the idea of reconstructing of human voice. Our basic aim will be to learn the voice of an existing voice sample, and then try to convert any given sound signal input, into that particular voice. For this we have reviewed some of the most recent work which emphasizes the most popular ways of voice reconstruction. In this regard henceforth, we will be enlisting an analysis part of the related work study and then we will propose our work.

System Design

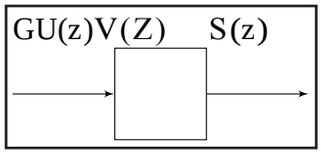
Linear Regression Techniques on a Z-Transform Implementation

This idea consists of two voices, source voice and target voice. The first sample i. e. the target voice is the one in which form we try to observe the required input. The source voice is the sample which contains the information that we need to have reconstructed (Raghunath *et al.,* 2013).

This methods implementation consists of three major stages, filter analysis, voice de-filtering and voice conversion. The broad outline of each of these methods is as follows. In the first stage, we second using Machine Learning techniques such as minimizing the mean squared error, the components unique to any human voice. Subsequently this is what we refer to as the human voice filter. In the second stage, we use speech signal processing techniques like Z-transform to get the segment of the speech from the given speech signal, by de-filtering the unique voice content of the particular human voice. In the third and final stage, we now pass this de-filtered voice into the human voice filter of the target voice, and obtain the final speech in the target voice (Raghunath *et al.,* 2013).

A rough idea of this is given below in the form of a block diagram. As shown in Figure 2, the central block is the filter V[z] which is a discrete time filter that models the human voice, and the filter Gu[Z] refers to the discrete time input which saves the words and other sounds in speech in some form. The output S[Z] consists of the exact speech samples recorded by us.

Figure 2: Voice Reconstruction Linear Regression Techniques



Auto-Regression on Stationary Time-Frames

The next basic idea is that of auto-regression on stationary time-frames where auto-regression is customized to the properties of the time-frame we consider. This is explained below. Auto-regressive techniques for voice conversion basically consist of three stages. At stage 1 we implement the Dynamic Time Warping; the second stage concentrates on K-Means clustering which emphasizes the fact that sound samples are stationary for relatively small time frames. This is justified by the fact that for small time frames, which are generally of 10ms, the

sound varies very less. Each of the frames would have auto-regressive techniques performed on them (Ye and Young, 2003).

Stage three is auto-regression for time-frame where we use the auto-regression means on relatively stationary frames. While auto-regression process proposes that output samples are dependent on a few previous output and input samples. This uses a feedback from output to determine the future output samples (Patel *et al.,* 2013).

Stage four and five consist of training phase and testing phase, respectively where in the training phase we will perform clustering and will obtain coefficients where as in the testing phase we will start once the source speaker's voice sample is obtained. Then we will first split it into stationary time samples, as in the training phase. These stationary time samples, initially in our testing phase, are then detected to be part of a cluster, among the set of clusters obtained in the training phase. The output in the next stage is achieved by use of the cluster obtained. The second stage consists of predicting the output frame given the cluster the input frame. Once the cluster has been obtained, we pull out the coefficients corresponding to the cluster, and use it to linearly generate the samples which mimic the output (Ye and Young, 2003).

Voice Construction Based on Pitch Synchronization Over-Lap Add (PSOLA) Algorithm

Our study says that in this approach, static voice conversion is basically taken care of. Static speech parameters are the parameters which cannot be changed by the speaker even on his willingness such as vocal tract structure modification, inherent natural pitch of speech etc. (Ganvit, Lokhandwala and Bhatt, 2012) It has got a very high utility in the multimedia application industry as it responds to the need for efficient storage of data. The quantitative analysis of the algorithm is dependent on the Quality factor (Q) and the Resemblance factor (R). The parameters are applied for diversified sample of voice.

Flow of Implementation of Psola (Mangayyagari and Sankar, 2007)

1. Application of silence removing algorithm for a silence free given input is the first step.
2. In the next step the voiced and unvoiced decision making algorithm takes care of the output of the previous step.
3. The voiced and unvoiced decision making algorithm is reutilized retrieving the qualitative data about the pitch.
4. The algorithm thus ends up right selection about the voiced and unvoiced aspect.
5. The qualitative data about the pitch is vital as markers for the pitch in the signal. The PSOLA algorithm assesses these indexes on the pitch for scaling down of each unique signal.
6. The flow process is two-fold; it scales the pitch and also converts the pitch. When we do have the data about the pitch indexes of both targets as well as the source segments then the mapping is performed for both and hence the conversion process continues. The conversion of the source to the target is achieved when the converted pitch indexes are processed through the PSOLA.

There are some other algorithms too like Linear Predictive Coding (LPC), Hybrid Harmonic/Stochastic (HYBRIDH/S) and last but not the least Time Domain-Pitch Synchronous Overlap Add (TD-PSOLA) which we are not discussing since they are not that beneficial in the present norms (Patel *et al.,* 2013).

Our Proposed Methodology

After analyzing a variety of approaches like Linear Transformations, Z-Transforms and the Continuous Probabilistic Transformation we arrived to a conclusion that all these methods take an assumption which is existence of good amount training data from both the source and the target speakers (Yamagishi *et al.,* 2012). The expectation of a parallel date so as to have a training data set consisting of the same textual segment spoken by both the source as well as the target speaker arises due to the application of least squared error estimation method.

Basically, a voice reconstruction module includes two parts, the initial training phase and the conversion process. In the first step of training procedure, the voice data from both the sources database of training set are parameterized and finally a function for conversion is generated to quire the similarity between the two. Any new speech can be transformed to a target voice segment

on the basis of the trained transformation function available. There are certain pressing requirements like the fact that the conversion function be almost lossless speaker transformation with very less noise, distortions and continuity snags in the modeled voice or speech. The above process may depend on a variety of factors but out of all of them abundance of training data for conversion function estimation is the most important one (Naniwa, Kondo and Kamiyama, 2012).

Most of the times the requirement for parallel training data can be fulfilled but there are arenas and applications which require voice modification for previously unknown voice data. Some of the prominent utility may be correction in the prosodic as well as the utterance correctness. The second utility may be reconstruction of voice for impaired speakers and also for anonymity of the speaker over phone (Connaghan and Patel, 2013). Our model is basically an extension and modification of the previous conventional approaches comprising of linear transformations. These transformations are also trained by the use of least square error methodology. A conventional recognizer on a HMM model may be utilized in the database where the target voice sample is to be kept parallel to the recognizer itself. The fresh arriving voice segment is marked in tandem to reach and retrieve the expected segments which can be mapped with the arriving voice. This process leads to the proper evaluation of the transformation functions.

The Intriguing Case of Constrained Training Data Set

Since we are considering the case where a constrained training dataset is available hence we need to find out the way to have the parameters for conversion. These parameters which are the transformation functions are now left to be retrieved from only one source and it is the source voice segment only. While doing so, the problem simplifies to that of tracing one or more frames of the target voice sample to amalgamate with subsequent voice sample of the unknown source. In the proposed system, the above task is accomplished by using a speech recognizer to index the target training voice data so that each new voice source can be used to retrieve similar frames from the target voice database. Much can be done on the speech recognizer part as this is at its primitive state. Thus retrieved voice samples may be used to predict the conversion functions by conventional methods (Patel *et al.,* 2011).

The step by step procedure or flow chart is as follows. Before the voice conversion of any fresh segment the available training data is managed in a database:

1. At the initial phase if the target sample data gets sequential of its own then its fine or else a conventional recognizer based on HMM model is used to get it done under all circumstances. Now each and every segment of the utterance in the target voice is assigned a particular number which is referred to as the state id.
2. Every training voice segment is properly evaluated and every Line Spectral Frequencies id is stored.

   Secondly, for each unknown utterance to be transformed which is the base criteria
3. The conventional recognizer based on HMM is utilized to monitor every new segment of speech so that the training dataset gets labeled. If the orthography of the utterance is known, forced alignment can be used. Otherwise, the utterance must be recognized. Finally every new speech segment is assigned with state id of the HMM model.
4. To trace the longest matching chain of the target vectors retrieved from the database they are first matched with the continuous chain of the state ids of the fresh speech segment. In the last phase of unit selection, an exact mapping of the corresponding source and target spectral vectors is performed.
5. Based on the mapping above, we compute a linear transformation as described later (Ye and Young, 2004).

The unit selection step as mentioned above is one of the most important steps. To achieve a guaranteed process that the continuous spectral evolution of the source is implemented on the transformed speech, it is essential to choose continuous target segments wherever applicable and possible. A constraint that helps achieve the longest matching state segments has been utilised. To explain and illustrate how this algorithm or process works let us consider the sequence of source state ids “1223445556”, subsequently upon close examination we can find that the longest matching series in the target data base is “12234” then the target spectral vectors corresponding to this subsequence are also extracted. This procedure then gets repeated with an expectation for a match for “5556” and hence forth until the total source sequence is symmetric. A concatenation process is initiated to get the final target vectors supplement the source one.

At this point it may be noted that the number of similar or parallel vectors which can be extracted from the voiced sounds in one utterance is generally not sufficient to train a robust transformation matrix hence, it is the training data which is to be changed by definition of the least squares criterion, the estimated matrix does provide, upto certain extent at least, a sufficient transformation of the training data. By using only the target vectors the output voice becomes machine like hence global transformations are better comparatively. The above issue may be addressed by transformation as the fresh vectors avoid discontinuities as they are already embedded in the source voice.

Genetic Algorithms

While dealing with the unit selection for training approach we thought of modifying the process and make it more convenient by adopting the Genetic Algorithm (GA) approach. Before we move on to principle of association of GA we would have a very brief overview of the actual GA and its requirements. Genetic algorithm has the following two requirements:

(i) A chromosome encoding of the solution domain.

(ii) A fitness function is to be generated to evaluate the solution domain.

The basic different phases of genetic algorithms are:

(i) Initialization: In this step a random selection of individual solutions is done to form the initial population.

(ii) Selection: A proportion of the existing population is selected to breed a new generation for every successive generation. The selection of individual solutions is made through a fitness-based process where fitter solutions (as determined by fitness function) have a greater chance of being selected than others.

(iii) Reproduction/ cross over and mutation.

(iv) Termination.

Some of the common termination conditions are:

(i) A solution satisfying the minimum criteria is obtained.

(ii) Fixed number of generation is reached.

(iii) The allocated resource has been consumed.

(iv) Highest ranking solution's fitness is reached or has reached a plateau such that successive iterations in longer lead to better results.

Now while choosing the continuous target segments wherever possible there may be certain GA based approaches for the same. While doing so we need to provide the basic ways of chromosome crossover and mutation. As a process what we generally do is while crossover we keep the initial 8 bits constants for both the parents then in the next 8 bits we apply the encoding. We have even enlisted certain modified encoding schemes as mentioned in the next section.

Table 1: Operators of Genetic Algorithm

Chromosome

Chromosome Name	*Encoding*
Chromosome 1	1100 0111 0101 1001
Chromosome 2	1001 1101 0111 0101

Crossover

Chromosome Name	*Encoding*	
Chromosome 1	1100 0111	0101 1001
Chromosome 2	1001 1101	0111 0101
Offspring 1	1100 0111	0111 0101
Offspring 2	1001 1101	0101 1001

Mutation

Chromosome Name	*Encoding*	
Offspring 1	1110 0111	0111 1101
Mutated 1	1100 0111	0111 0101
Offspring 2	1001 1001	0101 1001
Mutated 2	1001 1101	0101 1001

Encoding Schemes for Chromosomes

The basic genetic algorithms encoding techniques are Binary encoding, Permutation encoding and Value encoding. We have listed some relevant examples for each of the cases. What we want to do is to change the way of encoding which might be customized according to the need in the continuous spectral evolution of the source to precisely trace the matching series. The examples shown in Table 2 are just to show how those modifications may be shown.

Table 2: Crossover and Mutation on Various Encoding Schemes

Binary Encoding

Chromosome Name	*Encoding*		
Chromosome 1	110001	1101	011001
Chromosome 2	100111	0101	110101
Offspring 1	110001	1101	011001
Offspring 2	100111	0101	110101

Two point Crossover in Binary Encoding:-

Chromosome Name	*Encoding*			
Chromosome 1	1100	0111	0101	1001
Chromosome 2	1001	1101	0111	0101
Offspring 1	1101	1101	0101	0101
Offspring 2	1000	0111	0111	1001

Random Crossover

Use of Boolean Operation:-

Chromosome Name	*Encoding*			
Chromosome 1	1100	0111	0101	1001
Chromosome 2	1001	1101	0111	0101
Offspring 1 (bitwise AND)	1000	0101	0101	0001
Offspring 2(bitwise OR)	1101	1111	0111	1101

Arithmetic Crossover

Permutaion Encoding:-

Chromosome Name	*Encoding*
Chromosome 1	2537 81649
Chromosome 2	3781 95426
Offspring 1	2537 81946
Offspring 2	3781 25649

Value Encoding:-

Chromosomes	*Encoding*				
Chromosomes	2 . 3456	1 . 2145	5 . 2786	3 . 1023	0 . 7265
offspring	2 . 3456	1 . 1325	5 . 2786	3 . 1256	0 . 7265

Tampering with the Encoding, Crossover and Mutation Techniques

To formulate a Genetic Algorithm (GA) approach we can now conclude that the encoding cross over simulating plays the most important role. To list out the operators that impact it are:

a. Encoding Technique
b. Cross over point
c. Mutation

Now considering the fact that the major advantage of using GA is that they are broadly applicable and require little knowledge encoded in the system. However because of being knowledge poor approaches, GA fails to give satisfactory results and performance in some specific problems. Now our approach is to formulate a plan to remove this disadvantage due to knowledge poor condition by adopting the following approaches:

a. Making it partially knowledge rich- This aspect may be achieved by planning a proper encoding technique by either appending at the end or beginning a few bits or values which may be considered as constants so as to predict with greater accuracy the higher fitness values of the offspring.
b. Formulate an encoding plan which meets specific requirements-While performing encoding and crossover we may have an impact on the off springs by adopting a rule based cross over points.
c. A rule-based approach even in encoding techniques like permutation- Even in the case of permutation encoding we can follow certain rules.

Multiple Transforms

In our case the advantage is achieved by having a single global linear transform, but this use may lead averaging problem which may be avoided my having transforms in multiple numbers. By doing so, we have a better formant.

When there is fresh incoming voice then the use of interpolation method is the best way to have transforms of multiple nature (Naniwa, Kondo and Kamiyama, 2012). This is implemented in two phases, firstly the process of conversion of the source vectors are done by the use of a single global transform. Secondly the target components of the Gaussian Mixture Model's occurrence are computed and the posterior probabilities are then used as the interpolation weights.

Subsequently the multiple transform parameterizations can be done in the same way like conventional one. The constrained training data set initiates the practical problem acting as an hindrance to make out an estimate of the multiple transforms. But this data is just sufficient to train

a single dimensional linear transform matrix but not more than that. We will try to solve this problem by applying a more aggressive unit selection approach whereby each segment of the source voice is matched with many voice samples of the target database. To be more precise, the unit selection process is performed repeatedly until enough vectors for ex 300 number of GMM components have been extracted from the target database. At the end it can be inferred that the occurrence of the vector source is relatively high in the training set.

Conclusion and Future Work

While starting with the endeavour we decided to map the unique template of the "to be generated" voice segments for a singing tone but instead of that we tried to first critically analyze the prospect of reconstructing the basic phonetics. Hence in this work we have tried to formulate and propose a method of converting the speech of an source voice many a times it may be from unknown speaker to sound like that of some pre assigned target voice sample. The output may be further enhanced for better prosodic and formant structure. Lastly the prosodic and formant structure must be at par with the spectral transformation and this is very precisely another area needing endeavour.

References

Connaghan, K. P. & Patel, R. (2013) Impact of prosodic strategies on vowel intelligibility in childhood motor speech impairment. *Journal of Medical Speech Language Pathology*, 20(4), 133-139.

Ganvit,Y., Lokhandwala, M. A. & Bhatt, N. S. (2012). Implementation and overall performance evaluationof voice morphing based on PSOLA algorithm. *International Journal of Advanced Engineering Technology*, June, 3(2), 75-78.

Lemmetty, S. & Karjalainen, M. (1999). Review of Speech Synthesis Technology (Phd. Thesis)

Mangayyagari, S. & Sankar, R. (2007). *Pitch Conversion Based on Pitch Mark Mapping*. IEEE Proceedings Southeast Conference, 2007.

Naniwa, Y., Kondo, T. & Kamiyama, K. (2012). *Study on the Artificial Synthesis of Human Voice Using Radial Basis Function Networks*. In Proceedings in Information and Communications Technology, 4, 291-300.

Patel, R., Connaghan, K., Franco, D., Edsall, E., Forgit, D., Olsen, L., Ramage, L., Tyler, E. & Russell, S. (2013). The caterpillar: A novel reading passage for assessment of motor speech disorders. *American Journal of Speech Language Pathology*, February, 22, 1-9.

Patel, R., Hustad, K., Connaghan, K. P. & Furr, W. (2013). Relationship between prosody and intelligibility in children with Dysarthria. *Journal of Medical Speech Language Pathology*, 20(4), 95-99.

Patel, R., Niziolek, C., Reilly, K. & Guenther, F. (2011). Prosodic adaptations to pitch perturbation in running speech. *Journal of Speech Language and Hearing Research*, 54, 1051-1059.

Raghunath, A., Veerapandian, G. A. & Subramanian, V. G. (2013). Reconstruction of Human Voice for Impersonation. Final Report, November 18, 2013.

Yamagishi, J., Veaux, C., King, S. & Renals, S. (2012). Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. *Acoustical Science and Technology*, 33(1), 1-5.

Ye, H. & Young, S. (2004). *Voice Conversion for Unknown Speakers*. In Conference Proceedings of INTERSPEECH 2004-ICSLP. 8th International Conference on Spoken Language Processing.

Ye. H. & S. Young (2003). *Perceptually Weighted Linear Transformations for Voice Conversion.* In Proceeding of 8th European Conference on *Speech* and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland.