# FUNDAMENTAL FREQUENCY GENERATION FOR WHISPER-TO-AUDIBLE SPEECH CONVERSION

*M. Janke, M. Wand, T. Heistermann and T. Schultz*

Cognitive Systems Lab,
Karlsruhe Institute of Technology, Germany.

*K. Prahallad*

International Institute of Information
Technology, Hyderabad, India.

**ABSTRACT**

In this work, we address the issues involved in whisper-to-audible speech conversion. Spectral mapping techniques using Gaussian mixture models or Artificial Neural Networks borrowed from voice conversion have been applied to transform whisper spectral features to normally phonated audible speech. However, the modeling and generation of fundamental frequency ($F0$) and its contour in the converted speech is a major issue. Whispered speech does not contain explicit voicing characteristics and hence it is hard to derive a suitable $F0$, making it difficult to generate a natural prosody after conversion. Our work addresses the $F0$ modeling in whisper-to-speech conversion. We show that $F0$ contours can be derived from the mapped spectral vectors, which can be used for the synthesis of a speech signal. We also present a hybrid unit selection approach for whisper-to-speech conversion. Unit selection is performed on the spectral vectors, where $F0$ and its contour can be obtained as a byproduct without any additional modeling.

***Index Terms***— Silent speech interface, whisper-to-speech conversion, voice conversion, F0 generation.

## 1. INTRODUCTION

Speech is a result of time varying pulmonary excitation that is modified by the vocal tract. The excitation is typically followed by an oscillation of the vocal cords, which modulate the air flow expelled from the lungs and therefore may be closed or vibrate periodically. The period of this vibration is referred to as fundamental frequency ($F0$). Speech sounds are produced based on the type of excitation. While periodically vibrating vocal cords lead to voiced sounds, the lack of vibration leads to unvoiced sounds.

Whispered speech is produced when the vocal cords are adducted to produce a narrow constriction at the glottis. This results in excitation of vocal tract without the periodic vibration of the vocal cords. The articulation process itself remains the same as in normal audible speech. Whispered speech is of substantial interest both scientifically as well as for practical usage. From the speech production view, it is important to understand how a speaker embeds information into the speech signal without using $F0$. On the practical side, multiple application scenarios are imaginable: privacy in cell phone communication, no disturbance of bystanders or passing information only to the nearby are some evident examples. Another area of interest are patients with speech disabilities who are affected by an absent or handicapped pitch generation system and therefore can only produce whisper-like sounds.

All these problems are also tackled by the related area of Silent Speech Interfaces [1]. Silent Speech Interfaces have gained increasing interest over the last years. Most approaches like electromyographic speech recognition [2], ultra-sound based speech interfaces [3], and interfaces using Electromagnetic Articulography [4] aim at completely silently produced speech. A currently closer-to-market approach is the usage of Non-Audible Murmur (NAM) [5], which refers to whisper-like low amplitude sounds generated by laryngeal airflow noise and which is usually detected using a contact microphone attached to the skin. There also have been attempts to directly convert whispered speech to audible speech for a better understandability (e.g. [6] – [8]).

One general approach for this conversion is suggested by [9], where a statistical Gaussian mixture model is trained to predict both spectral features and $F0$ of audible speech from the spectral vectors of whispered speech. However, from the studies in [10] and [11], it is understood that $F0$ is encapsulated in whispered speech in a non-trivial way. Hence, instead of using the input whispered speech, we investigate whether spectral vectors of audible speech obtained as a result of the statistical transformation from the whispered speech bear a better correlation with fundamental frequency.

The remainder of this paper is organized as follows. Section 2 provides the details on the database used for our experiments, followed by a study of the general necessity for a whispered-to-audible speech conversion in section 3. Section 4 explains the proposed baseline system, and differentiates this approach from other techniques. Section 5 discusses a hybrid unit selection approach for the conversion of whisper-to-audible speech to enable a more natural $F0$ generation, which is followed by a listening test evaluation in section 6. The work is concluded with section 7.

## 2. DATABASE INFORMATION

We use a dataset with normal and whispered read speech of five persons (four male, one female). While the mother tongue of all subjects is non-English, their English pronunciation skills range from good to very good. Their age ranges from 21 to 33 years. Each person recorded 200 phonetically balanced English sentences per speaking mode, i.e. a total of 400 sentences, which originated from the broadcast news domain. All sentences were recorded using a smartphone microphone with a sampling rate of 16 kHz once in normal voice and once in whispered voice in random order. Since one of the applications for whispered speech is privacy in phone communication we used a smartphone microphone instead of a standard headset microphone. All recorded audible and whispered speech utterances are divided into three sets for training (70%), development (15%), and evaluation (15%) per speaker. Table 1 gives detailed information about the durations of the recorded utterances.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

**Table 1**. Data corpus information for recorded audible (aud) and whispered (whis) utterances, including speaker breakdown.

| Speaker | Average data length, in [s] for aud/whis | | | Total amount of data (mm:ss) |
|---|---|---|---|---|
| | Train | Dev | Eval | |
| 1 | 533/513 | 109/104 | 113/104 | 12:36/12:02 |
| 2 | 554/596 | 110/121 | 116/125 | 13:01/14:01 |
| 3 | 514/541 | 102/106 | 110/116 | 12:06/12:43 |
| 4 | 532/530 | 107/107 | 112/112 | 12:32/12:29 |
| 5 | 478/497 | 102/105 | 108/114 | 11:28/11:56 |
| Total | 2611/2677 | 530/543 | 560/571 | 61:42/63:11 |

## 3. PREFERABILITY OF WHISPERED SPEECH VERSUS NORMAL SPEECH

Since whispered speech is typically perceived well by human beings, it is natural to question the necessity of a whisper-to-audible speech conversion. Given natural recordings of both audible and whispered speech, we investigate whether listeners prefer whispered or audible speech in quiet office environments and in noisy conditions. For testing, we use 36 randomly selected English sentences from our data corpus described in section 2.

The TestVox [12] framework is used to conduct AB preference listening tests. A set of 10 subjects with non-English mother tongue take part in the listening tests. For all 36 test sentences, each participant listens to the whispered and normally voiced version and is asked which version he/she prefers considering the intelligibility. A third, neutral option is also given. To minimize bias, we randomize both the order of the 36 test sentences, as well as whether the whispered or the normal utterance is played first. Every participant listens to one half of the sentences via laptop speakers and the other half of the sentences using a clip-on-headphone on his preferred ear (to simulate the act of hearing on a cell phone). In order to examine the effect of a loud environment to the intelligibility of whispered speech, five of the ten listeners carry out the tests in a quiet office environment. The remaining five participants conduct the tests at a cafeteria during lunchtime, a particularly loud and noisy environment. All tests are supervised by an instructor to guarantee that all participants follow the same rules.

Table 2 shows the results of the listening tests and states that whispered speech is preferred substantially less (about $4\%$) even in the quiet office environment. We also observe that there is no preference in about $39\%$ to $44\%$ of the utterances. We assume this is an equal bias towards audible and whispered speech. Audible speech is preferred in $51\%$ to $59\%$ for all situations. These listening tests suggest the strong preference of the listeners towards audible speech and hence justify the efforts of performing a transformation from whisper-to-audible speech.

**Table 2**. Preference in percentages for audible speech (Aud), whispered speech (Whis), or neither. Each test condition has five listeners rating a total of 180 sentences.

| Test-cond. | Aud | Whis | Neither |
|---|---|---|---|
| Quiet environment: Laptop | 51.1% | 4.4% | 44.4% |
| Quiet environment: Headphone | 54.4% | 4.4% | 41.1% |
| Noisy environment: Laptop | 55.6% | 3.3% | 41.1% |
| Noisy environment: Headphone | 58.9% | 2.2% | 38.9% |

## 4. WHISPER-TO-AUDIBLE SPEECH CONVERSION

### 4.1. Feature extraction and alignment

To extract features from the speech signal, an excitation-filter model of speech is applied. 25 Mel-cepstral coefficients (MCEPs) [13] are extracted as filter parameters and fundamental frequency ($F0$) estimates are derived as excitation features for every 5 ms. As the durations of the whispered and normal utterances typically differ (see table 1), dynamic time warping is used to align MCEP vectors of whispered and audible speech.

### 4.2. Whisper-to-audible mapping

In this work, we use an artificial neural network (ANN) to perform the mapping from spectral vectors of whispered speech to audible speech. However, such mapping can also be performed using Gaussian mixture models [9] and we refer to this component abstractly as mapper. Details about the training and topology of this ANN mapping can be found in [14].

To perform the spectral mapping of whispered speech to audible speech, the 25 dimensional MCEPs from whispered speech are concatenated with their $\Delta$ and $\Delta\Delta$ features to build a 75 dimensional vector, which was mapped to a 25 dimensional vector of audible speech using the ANN model topology $75L\ 90N\ 90N\ 25L$. Here $L$ represents a linear activation function, $N$ represents the nonlinear $tanh$-function and the numbers refer to the amount of neurons used in each layer. The structure of the ANN is chosen on empirical basis from prior experiments.

Figure 1(a) shows the general approach presented in [9]. As can be observed, the mapper is trained to predict both spectral features and $F0$ of audible speech from the spectral vectors of whispered speech. We therefore refer to this $F0$ mapping approach as $WhisToF0$. However, from the studies in [10] and [11], it is understood that $F0$ is encapsulated in whispered speech in a non-trivial way. Hence, we want to investigate whether spectral vectors of audible speech obtained as a result of the statistical transformation from the whispered speech can bear a better correlation with fundamental frequency.

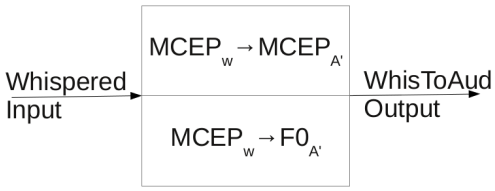
To study this correlation, we use a linear regression fit to predict $F0$ from the first cepstral coefficient $C0$ of whispered and audible speech. Table 3 shows the mean square error (MSE) in predicting $F0$ from $C0$ of whispered, audible and whisper-to-audible mapped speech. The MSE in predicting $F0$ using whisper $C0$ exceeds the MSE when using $C0$ from audible and whisper-to-audible $C0$. At the same time, the MSE in predicting $F0$ using $C0$ of whisper-to-audible is smaller than using $C0$ from whisper.

**Table 3**. Mean square error (MSE) of the linear regression fit on the first cepstral coefficient ($C0$) of whisper, audible and whisper-to-audible mapped spectral vectors.

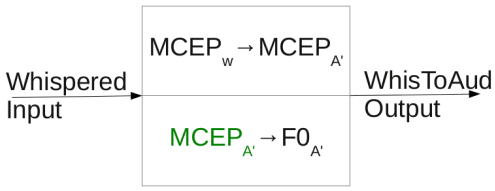
| Type of $C0$ | MSE ($Hz^2$) of linear regression fit of $C0$ and $F0$ |
|---|---|
| Whisper | 772 |
| Audible | 710 |
| Whis-to-aud (mapped) | 721 |

### 4.3. $F0$ modeling

From table 3, it can be inferred that spectral vectors of audible and whisper-to-audible speech can be used for predicting $F0$ and its




(a) $WhisToF0$ Conversion Approach




(b) Proposed $MappedToF0$ Conversion Approach

**Fig. 1**. *Architecture of (a) whisper-to-audible speech conversion approach followed by [9] ($WhisToF0$) and of (b) proposed whisper-to-audible speech conversion ($MappedToF0$). MCEPs represent spectral features, and $F0$ represents fundamental frequency. W refers to whispered speech features, and A' refers to whisper-to-audible converted features.*

contour. Regarding this result, we thus propose the architecture as shown in figure 1 (b) for modeling $F0$ and its contour in the case of whisper-to-audible speech conversion. Since the whisper-to-audible converted MCEP$_{A'}$ features instead of the whispered MCEP$_W$ are taken as input for the $F0$ mapping, we refer to this approach as $MappedToF0$.

Following this approach, we model $F0$ and its contour as a non-linear mapping of spectral vectors produced from the output of the whisper-to-audible spectral mapper. Once the spectral mapper (referred to as ANN–I) is trained with the time-aligned MCEP data (using Dynamic Time Warp) of whispered and audible features, the entire training data is fed as input to obtain the predicted audible MCEPs at the output layer.

The predicted MCEPs (MCEP$_{A'}$) are used to train another ANN model (referred to as ANN-II), whose output is a 51 dimensional vector representing the $F0$ contour. Instead of predicting a single $F0$ value, we train the ANN-II to model a $F0$ contour of about 250 ms (i.e., appending $F0$ values from the left and right context of 25 frames). The choice of 250 ms is made in view of the average syllable length [15].

Thus for every 5 ms, we obtain a 51 dimensional $F0$ vector. The final $F0$ contour is obtained by picking the middle value of this vector and performing a mean smoothing. Figure 2 shows the $F0$ contour predicted from the mapped MCEPs (i.e., from the output of ANN-I) of the whispered speech. The predicted $F0$ demonstrates that mapped MCEPs are useful features to predict $F0$ and its contour. The topology of ANN-II used in these experiments is $25L$ $75N$ $12N$ $75N$ $51N$. The middle layer represents a compression layer, which projects the data onto a lower dimensional space, for extracting compact representations.

To finally synthesize a whispered-to-audible converted speech signal we use the Mel Log Spectrum Approximation (MLSA) filter method [16], which takes the generated $F0$ and the mapped MCEPs as input.

## 5. A HYBRID FRAMEWORK

In this section, we describe our experiments on selecting a unit (either in the parametric space or directly from the wavefiles) for the purpose of whisper-to-audible speech conversion. This idea is motivated from a hybrid unit selection framework in text-to-speech systems [17]. Given that we have target spectral vectors (i.e. MCEPs predicted from ANN-I), one could use these vectors to search for a suitable unit in the training set. This enables to pick a nearest neighbor MCEP vector and the corresponding $F0$ from the training set. One could also opt to use the speech segment directly to avoid signal processing. The steps followed in our approach are as follows:

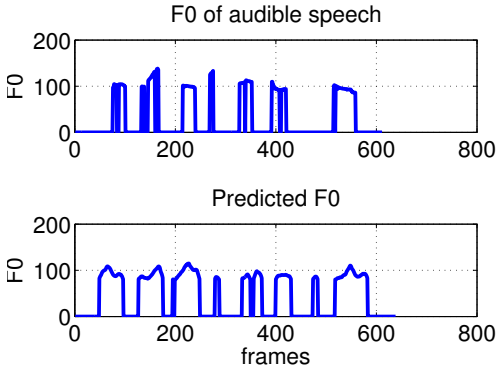



**Fig. 2**. *$F0$ of audible speech and predicted from whis-to-aud mapped MCEPs. (No time alignment used)*

- During training, the set of MCEP vectors from the training data are used to build a dictionary. Each entry in the dictionary is a 125 dimensional vector, i.e., a joint MCEP vector with a left and right context of 2 frames. These entries are indexed along with the corresponding $F0$ values and with their time stamps in the corresponding wave files.
- The output vectors predicted by ANN-I on the development set are appended with the left and right context information and the dictionary is searched for their nearest neighbors. While one could use a sophisticated search algorithm, as an initial step, we choose the nearest neighbor approach.

  Subsequently, the MCEPs and the F0 of the corresponding nearest neighbor are used to synthesize the speech signal via the MLSA filter. This approach is referred to as $synth1$.
- As another method, the corresponding time stamps are used to slice the unprocessed speech waveform segments and concatenate them using overlap and add method [18]. This approach is referred to as $synth2$.

To evaluate the effectiveness of the proposed mapping techniques, Mel-cepstral distortion (MCD) [19] is computed between MCEPs of audible and converted whisper-to-audible speech after aligning them using dynamic time warp. The MCD scores are shown in table 4. While there is only little difference between the hybrid unit selection approaches $synth1$ and $synth2$, MCD scores achieved with ANN are clearly lower. Additionally, we compute the distortion of the 0th Mel coefficient, which represents the signal's energy and is therefore ignored in speech recognition appli-

cations. Since we want to investigate the prosodic properties of the whisper-to-audible mapping, we evaluate this coefficient separately in figure 3. It is noticeable that $synth1$ achieves the smallest distortion, implying that this may be related to an improved prosodic modeling.

**Table 4**. Mel-cepstral distortion (MCD) and standard deviation for whisper-to-audible mapping using ANN model (ANN-MCD) and our proposed hybrid approaches ($synth1$ and $synth2$). The MCD scores (lower is better) are computed on the development set of 30 wavefiles.

| Spk | ANN MCD | Synth1 MCD | Synth2 MCD |
|---|---|---|---|
| 1 | 4.34 (0.24) | 5.18 (0.28) | 5.21 (0.28) |
| 2 | 4.77 (0.28) | 5.65 (0.31) | 5.64 (0.29) |
| 3 | 5.59 (0.34) | 6.44 (0.35) | 6.43 (0.34) |
| 4 | 5.16 (0.39) | 6.00 (0.38) | 5.99 (0.36) |
| 5 | 5.58 (0.30) | 6.55 (0.32) | 6.81 (0.27) |

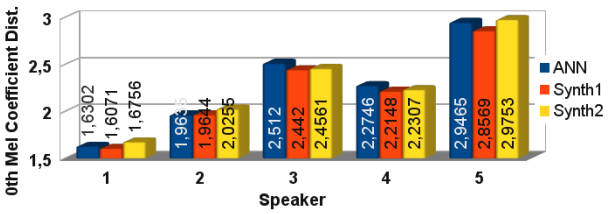



**Fig. 3**. *Distortion (logarithmic scale) of* $0th$ *Mel coefficient for whisper-to-audible mapping using ANN model and hybrid approaches* $synth1$ *and* $synth2$.

Figure 4 shows the spectrogram of the recorded audible speech and the converted whisper-to-audible speech from ANN (i.e., ANN-I + ANN-II), $synth1$ and $synth2$ methods. It is noticeable that $synth2$ produces stopping sounds due to the concatenation of the audible sound units, which can be easily seen in the spectrogram. On a closer look (e.g. between $0.8s$ and $1s$) it can be observed that the formants are better modeled with $synth1$ and $synth2$, regarding a flat contour with the ANN approach.

## 6. LISTENING TEST EVALUATION

Since MCD scores do not consider the quality of the $F0$ modeling, we evaluate our mapping approaches in conducting two AB preference listening tests, which again include a third, neutral option.

The first test is designed to evaluate the output of $synth1$ versus $synth2$, the second listening test aims to distinguish between $synth1$ and the $WhisToF0$ ANN approach. Each participant listens to both mapping outputs and has to give a preference considering the naturalness. We intentionally decided to ask for naturalness instead of intelligibility, since this is more affected from a proper $F0$ contour. To minimize bias, we again randomize the order of test sentences and both output files. 17 listeners participate in the $synth1$ versus $synth2$ test (total of 425 utterance-pairs) and there are 14 participants in ANN versus $synth1$ test (total of 700 utterance-pairs).

Table 5 shows both listening test results. $Synth1$ gives significantly ($p = 0.0293$) better results than $synth2$, while a comparison of $synth1$ with the $WhisToF0$ ANN mapping shows no significant preference.

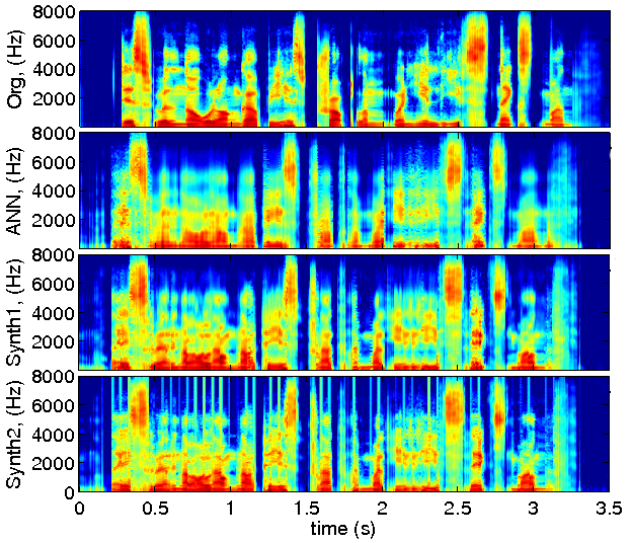



**Fig. 4**. *Spectrogram of audible speech (org), output from ANN,* $synth1$ *and* $synth2$ *methods (no time alignment used) of the utterance "This will no longer be his problem when he leaves next month".*

Having a detailed look at the mapped utterances depicts two drawbacks of the compared methods: ANN gives a smooth output with a flat prosodic contour that sounds very robotic, while $synth1$ gives a more variable prosodic contour with unnatural stopping sounds due to the alignment of the units. It is therefore hard for the listener to estimate a preference considering a broad term like naturalness.

**Table 5**. Preference in percentages for $synth1$ (A) versus $synth2$ (B) on upper half, and hybrid approach $synth1$ (A) versus ANN approach (A) or neither.

| Test-cond. | A | B | Neither |
|---|---|---|---|
| $synth1$ vs $synth2$ | 42.35% | 33.18% | 24.47% |
| $synth1$ vs ANN | 34.29% | 37.57% | 28.15% |

## 7. CONCLUSION AND FUTURE WORK

To emphasize the usefulness of a whisper-to-speech transformation, we conducted a listening test which shows the preference of the listeners towards audible speech. In order to improve the modeling of $F0$, we successfully used the mapped spectral vectors instead of the whispered input spectral vectors.

Additionally, we showed that a hybrid unit selection framework can be adapted for whisper-to-speech conversion. This approach was performed on the spectral vectors, where $F0$ and its contour could be obtained without any additional modeling and thus gives the opportunity to generate a natural $F0$ contour, which is hard to obtain by considering whispered speech only. This approach reduced the distortion of the $0th$ Mel coefficient and is a first step towards a natural $F0$ generation.

Since this is a preliminary study, there are several factors (e.g. stopping sounds due to unit concatenation) that will be optimized in the near future.

## 8. REFERENCES


[1] Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., and Brumberg, J. S., "Silent Speech Interfaces", Speech Communication, 52(4), pp. 270–287, 2010.

[2] Maier-Hein, L., Metze, F., Schultz, T., Waibel, A., "Session Independent Non-Audible Speech Recognition using Surface Electromyography", IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 331–336, 2005.

[3] Denby, B., Stone, M., "Speech Synthesis from Real Time Ultrasound Images of the Tongue", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. I685–I688, 2004

[4] Fagan, M.J., Ell S.R., Gilbert J.M., Sarrazin E. and Chapman P.M., "Development of a (Silent) Speech Recognition System for Patients following Laryngectomy", Medical Engineering and Physics, 30 (4), pp. 419–425, 2008

[5] Nakajima, Y., Kashioka, H., Shikano, K., Campbell, N., "Non-Audible Murmur Recognition Input Interface using Stethoscopic Microphone Attached to the Skin. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 708–711, 2003.

[6] Tran, V. A., Bailly, G., Loevenbruck, H., and Toda, T. "Predicting F0 and Voicing from NAM-Captured Whispered Speech", Proc. Speech Prosody, pp. 133–136, 2008.

[7] Sharifzadeh, H. R., McLoughlin, I. V., and Ahamdi, F., "Voiced Speech from Whispers for Post-Laryngectomised Patients", IAENG International Journal of Computer Science, 36(4), pp. 367–377, 2009.

[8] Passos, A. P., "A Lightweight Processing for Conversion of Whispering Voice into Normal Speech", International Conferenceon Audio Language and Image Processing (ICALIP), pp. 74–79, 2010.

[9] Toda, T., Shikano, K., "NAM-to-Speech Conversion with Gaussian Mixture Models", 9th European Conference on Speech Communication and Technology (Interspeech), pp. 1957–1960, 2005.

[10] Meyer-Eppler, W. "Realization of Prosodic Features in Whispered Speech." The Journal of the Acoustical Society of America 29, pp. 104–106, 1957.

[11] Itoh, T., Takeda, K., and Itakura, F. "Acoustic Analysis and Recognition of Whispered Speech", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 389–293, 2002.

[12] Parlikar, A., "TestVox: Web-based Framework for Subjective Evaluation of Speech Synthesis", Opensource Software, https://bitbucket.org/happyalu/testvox, 2012.

[13] T. Fukada, Tokuda, K., Kobayashi, T., and Imai, S., "An Adaptive Algorithm for Mel-Cepstral Analysis of Speech," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 137–140, 1992.

[14] Desai, S., Raghavendra, E. V., Yegnanarayana, B., Black, A. W., and Prahallad, K., "Voice Conversion using Artificial Neural Networks", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3893–3896, 2009

[15] Greenberg, S., Carvey, H., Hitchcock, L., and Chang, S., "Temporal Properties of Spontaneous Speech - A Syllable-Centric Perspective", Journal of Phonetics, 31(3), pp. 465–485, 2003.

[16] Imai, S., "Cepstral Analysis Synthesis on the Mel Frequency Scale", IEEE International Conference on Acoustics, Speech, and Signal Processing Vol. 8, pp. 93–96, 1983.

[17] Black, A. W., Bennett, C. L., Blanchard, B. C., Kominek, J., Langner, B., Prahallad, K., and Toth, A., "CMU Blizzard 2007: A Hybrid Acoustic Unit Selection System from Statistically Predicted Parameters", Blizzard Challenge Workshop, Bonn, Germany, 2007.

[18] Verhelst, W., "Overlap-Add Methods for Time-Scaling of Speech", Speech Communication, 30(4), pp. 207–221, 2000.

[19] Toda, T., Black, A. W., and Tokuda, K., "Mapping from Articulatory Movements to Vocal Tract Spectrum with Gaussian Mixture Model for Articulatory Speech Synthesis", Fifth ISCA Workshop on Speech Synthesis, pp. 31–36, 2004