A hybrid harmonics-and-bursts modelling approach to speech synthesis

Jonas Beskow[1] *and Harald Berthelsen*[2]

[1]KTH Speech, Music and Hearing.
100 44 Stockholm, Sweden
[2] STTS Speech Technology Services
Stockholm, Sweden

beskow@kth.se, harald@stts.com

Abstract

Statistical speech synthesis systems rely on a parametric speech generation model, typically some sort of vocoder. Vocoders are great for voiced speech because they offer independent control over voice source (e.g. pitch) and vocal tract filter (e.g. vowel quality) through control parameters that typically vary smoothly in time and lend themselves well to statistical modelling. Voiceless sounds and transients such as plosives and fricatives on the other hand exhibit fundamentally different spectro-temporal behaviour. Here the benefits of the vocoder are not as clear. In this paper, we investigate a hybrid approach to modeling the speech signal, where speech is decomposed into an harmonic part and a noise burst part through spectrogram kernel filtering. The harmonic part is modeled using vocoder and statistical parameter generation, while the burst part is modeled by concatenation. The two channels are then mixed together to form the final synthesized waveform. The proposed method was compared against a state of the art statistical speech synthesis system (HTS 2.3) in a perceptual evaluation, which reveled that the harmonics plus bursts method was perceived as significantly more natural than the purely statistical variant.

Index Terms: hybrid speech synthesis, statistical speech synthesis, concatenation, spectro-temporal filtering, crowd source evaluation.

1. Introduction

Today, there are two dominating approaches to generating synthetic speech: concatenative methods and parametric methods. The concatenative methods essentially copy entire waveform segments retain all the fine spectral and temporal detail of the original speech and have made it possible to synthesize highly natural and intelligible speech, but suffer from lack of flexibility and extensive data requirements and the notorious problem of making two segments of speech recorded at different times and in different context sound as if they belong together.

Parametric approaches typically rely on the principles of source-filter decomposition of the speech signal [1]. The source-filter model is a useful way of decomposing the speech signal because the source and the filter typically convey different and complementary information, and the source-filter decomposition allows independent control of fundamental frequency and various voice source characteristics – which often are of supra-segmental relevance - and the vocal tract characteristics, which govern vowel quality. The above is true primarily for voiced parts of the speech signal. The source-filter model will handle voiceless speech as well, but there the advantages of the model are not so obvious. As Fant argues, any transient noise burst can be modeled using a noise source and a time-varying filter. In this case, the noise source can be thought of as turbulent airflow in a constriction in the vocal tract, and the filter will correspond to the part of the vocal tract above the constriction. Early formant synthesizers included separate filter branches for voiced sounds and frication sounds [2][3], but the typical way this is implemented in todays vocoders used in parametric speech synthesizers, is using a single filter stage which alternates between the roles of a vocal (formant) filter – which typically models smooth and continuous spectral variations - and a filter that shapes frication noises and transient explosion bursts which operate on a much finer time scale.

In this paper, we investigate a hybrid approach to modeling the speech signal, where we first separate the speech into an harmonic part and a noise burst part. Then we use a vocoder to model the harmonic part, while the noise bursts are stored in a library as raw waveforms that can be overlaid on the vocoded harmonics at synthesis time. The remainder of the paper is organized as follows: first we look at related work, then we describe the separation stage where the speech signal is divided into harmonics and bursts. After that we describe our synthesis procedure, followed by a perceptual evaluation where the proposed approach is compared against a state-of-the-art parametric speech synthesis system. Finally we draw

conclusions and point out directions where more research is needed.

2. Related work

The idea of combining parametric and concatenative approaches to synthesis in order to reap the benefits of the two methods and circumvent some of the shortcomings is not new.

Several systems have been described that combine statistical parametric speech synthesis and unit selection approaches in order to optimize target cost [5][6][7]. [8] described a hybrid between statistical and concatenative synthesis where a statistical parametric system is used to generate the speech from an HMM system, but a concatenative unit selection system is executed in parallel and used to improve the quality of the statistical system. This is done in an effort to combine the robustness of the HMMs with the naturalness of concatenated units.

Another class of hybrid systems, more close to what we propose here, are those that acutally mix natural speech segments with those generated from parametric models.
An early example of such an approach is [4]. This system incorporated sampled voiceless consonants in a rule-based formant synthesizer. While similar in spirit to the work in the present paper, the rule-based approach makes the methods involved very different. For example, consonant waveforms had to be manually analyzed and segmented into units to be used in the synthesizer.
More recently, [9] proposed a system that combines statistical parametric synthesis and concatenation. In their system, they employ a hybrid dynamic-path algorithm that allocates natural segments along with statistical boundary-constrained model-generated segments, and a corresponding hybrid speech feature-vector generating algorithm.

A similar approach is taken by [10], who describes a multi-form synthesis engine capable of combining natural speech segments and segments generated from statistically generated parameter trajectories. An interesting feature of this system is that it is possible to specify a Model-Template-Ratio that dictates the probability of a speech segment being generated by statistical models rather than taken from natural speech (template). There are many differences between these designs and the one proposed in the present paper, but one critical one is that all of these approaches mainly work by splicing and concatenating segments in time, whereas our approach models two *channels* with different spectro-temporal properties and overlays them to form the final speech wave.

Yet another type of hybrid system is proposed in [11]. In this case, a library of sampled waveforms representing glottal pulses are used to excite the filter in a statistical speech synthesis system. This approach has certain similarities (and also gave some inspiration) to the method described in the current study. One such similarity is that the units of concatenation in both methods are of transient nature (albeit on a different time scale). This fact alleviates many of the problems traditionally encountered in speech unit concatenation such as seamlessly joining segments together. Another similarity is that spectral continuity over time in both methods is ensured by statistically modeled parameters.

3. Harmonics and burst separation

The basic idea in this contribution is that we decompose the speech signal into two components, or channels, with different spectro-temporal properties and then use different techniques to model and these two components in the synthesizer. Finally we recombine the two channels to retrieve the complete synthesized waveform.

The *harmonic* part of the speech signal, $X_h(n)$ stems from the vibrations of the vocal folds, filtered through the vocal tract. This is a quasi-periodic signal that will exhibit self-similarity over time. Comparing two spectrum sections taken closely in time, the difference is expected to be small. The other part of the speech signal that will be extracted will be referred to as the *burst* component and will be denoted $X_b(n)$. This signal stems from frication noises due to turbulent airflow in the vocal tract, both at narrow constrictions (fricative sounds) and from the release of pressure build-up at occlusions in the vocal tract (plosive sounds). Looking at the spectro-temporal patterns of this type of sounds, we don't expect two consecutive spectrum sections to show much similarity. On the other hand, given the aperiodic nature of the bursts, and the fact that they haven't been filtered through the full length of the vocal tract, we expect energy to be more distributed along the frequency axis than in the harmonic component, where it is concentrated to a number of discrete frequencies.

The first step of the separation is to decompose the signal $X(n)$ into a harmonic component $X_h(n)$ and an inharmonic component $X_i(n)$. To do this we employ the spectrogram filtering approach described by [12] and the MATLAB implementation provided by the authors. The basic idea of the method is to compute a spectrogram of the input signal, and apply median filtering with different kernels to separate the signal into components with different spectro-temporal properties.

For the spectrogram, we use a window length of 10 ms and an 85% overlap between frames, leading to a 1.5 ms time shift between consecutive frames and a frequency bin size of 100 Hz.
To extract the harmonic part $X_h(n)$, we use a kernel of 100 ms (67 steps) in the time direction and 100 Hz (1 step) in the frequency direction. For the inharmonic part $X_i(n)$, we used a vertical kernel with 8000 Hz (80 steps) in the frequency direction and 1.5 ms (1 step) in the time direction.
We used only one filtering pass. The method proposed by [12] also allows for iterative re-estimation of the components but in our experiments we got best results with one pass filtering.

After the initial filtering step, $X_i(n)$ will mainly consist of noise bursts associated with frications and plosive consonants, as well as certain periodic content stemming from irregularities in the voice source pulses. Since we wish to be able to model the burst part of the signal without taking pitch into account, our next step is to silence any portions of the burst signal that contain periodic information. We use the *REAPER F0* tracker [13] to calculate a binary voicing decision for the filtered burst signals tracks with a 200 Hz frame rate, and then we calculate the burst signal $X_b(n)$ as

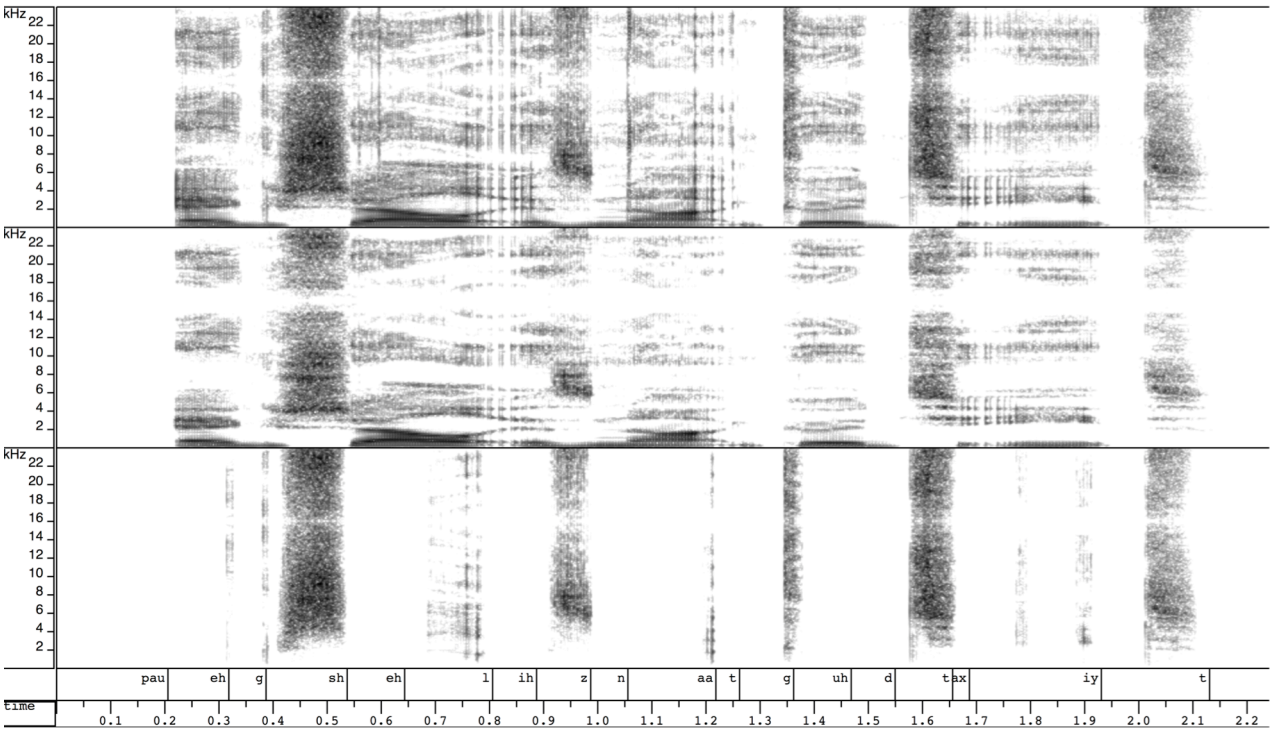

Figure 1: Separation of speech into harmonic and burst components. Top spectrogram shows the original signal X(n) for the sentence Egg shell is not good to eat. The middle pane shows the harmonic component $X_h(n)$, while the bottom pane shows the bursts component $X_b(n)$.

$$X_b(n) = (1\text{-}X_v(n))\, X_i(n)$$

where $X_v(n)$ is the binary voicing decision interpolated to the sampling rate (48 kHz).
Figure 1 shows the spectrograms of the original signal $X(n)$, the harmonic signal $X_h(n)$ and the burst signal $X_b(n)$.

4. A Burst Library

The extracted bursts are organized into a burst library in order to facilitate synthesis. We used 593 sentences from the *US SLT* voice from the CMU Arctic database [14]. The $X_b(n)$ signal was calculated as described above and segmenting into it isolated noise bursts. This is done by calculating an energy envelope as the squared signal smoothed by a moving average of 100 samples (approx. 21 ms). Any part of the signal where the energy envelope stays above a certain threshold e_{min} for a duration of t_{min} is considered an isolated burst segment. The thresholds were adjusted manually. All of the other steps in building the library are fully automatic.

4.1. Burstables

Next, the identified noise bursts are associated with phones in the database. For convenience and for the purposes of this study we will define a set of phonemes that (usually) are associated with distinct noise bursts. We will call this set of phonemes *burstables* and in this study they were defined as the following set of fricatives and voiceless plosives: /p/, /t/, /k/, /f/, /s/, /θ/, /z/, /ʒ/, /ʃ/, /tʃ/ and /dʒ/.
For every *burstable* in the *CMU Arctic* sentences, it was checked whether the onset of a noise burst fell within the boundaries of the phone. If so, this noise burst was added to the library, together with the following information:

- Phoneme (one of the ones listed above)
- Preceding phone
- Following phone
- Duration of the phone
- Onset time of the burst relative to phoneme duration
- Duration of the burst
- Energy of the burst in dB

This information is stored, together with the actual waveform data for the burst.
For 67% of the *burstable* phones a burst was detected and added to the library, which resulted in 2996 entries. In order to improve consistency of the automatically extracted noise bursts, a basic pruning of outliers was done by removing all bursts falling in the lowest and highest 1% with respect to *duration*, *energy* and *onset*. This left 2782 entries in the library.

5. Hybrid Synthesis Procedure

We now turn to describe how we use our burst library in combination with the HTS parametric speech synthesis system [15] to synthesize speech.

5.1. Synthesizing the harmonic track

We use the procedure described in sect. 3 to extract the filtered harmonic signal $X_h(n)$ from the *US SLT ARCTIC* database with a sampling frequency of 48 kHz. Then we use the scripts from the *speaker dependent training demo (English)* of the HTS system ver *2.3* to train a HTS voice on the harmonic component. This voice can then be used together with *hts_engine* to synthesize the harmonic component of the speech. Note that since we will be synthesizing and overlaying two components, we will retrieve the segment durations from the synthesis of the harmonic component and use that when synthesizing the burst component.

It is interesting to compare the properties of this voice to the voice trained on raw data (i.e. the standard HTS demo voice). The *F0* contours will be slightly different and the de-voicing that sometimes happens in HTS voices seems to occur less frequently. We believe that this may have something to do with the fact that the filtered harmonic signal is smoother with less transients than the raw speech signal and thus less prone to introduce errors in *F0* tracking or estimation of the vocal tract filter function. But this is a matter that requires a separate investigation.

5.2. Synthesizing the burst track

Given the burst library described in sect 4, as well as a phonetic specification, including segment durations, of the speech to be synthesized, we are faced with the tasks of constructing a burst component that, when overlaid on top of the harmonic component, will restore all the transient and fricative components that were filtered out earlier.

Our basic approach is as follows: for each phone in the utterance, first determine if the phone should have a burst, and if so, determine onset time and find the most appropriate entry in the burst library given the synthesis context., and simply copy the corresponding waveform into the r place.

5.2.1. Energy class

The first thing to note here is that not every instance of the phonemes in the list of *burstables* (see 4.1) will actually exhibit a burst. During construction of the burst library, almost 30% of these phones were not associated with a burst (or there were been bursts that fell below the energy or minimum time threshold). This implies that we need a way to predict whether or not there should be a burst at all in a given phone slot. We can also note that there is quite some variability in terms of intensity of the noise bursts. So if we decide that there should be a burst, we must also know how strong it should be.

We give each burst in the library an energy class rating 1,2 or 3 based on it's energy value, by looking at which percentile it falls into (class 1: <33%, class 2: 33-67% or class 3: >67%.) on a per-phoneme basis. We trained a regression tree was using the *Wagon* tool [16] to predict energy class on a continuous scale from 0 – 3 where 0 corresponds to no burst. The input features used in the prediction were *phone duration* plus *all the features used in the HTS demo system full context labels* (i.e. phone- syllable, phrase and utterance level features), amounting to a total of 54 features.

5.2.2. Burst onset time

Another important aspect is the timing of the burst. The onset time in relation to the start of the phone will vary depending on several factors such as the phoneme class (fricatives will have an onset close to the beginning of the phone while plosives will have the onset closer to the middle or second half), the context and the duration. To account for this variability we trained a regression tree to predict the onset time for the burst relative to the phone boundaries (0 - 1 where 0 corresponds to the start of phone and 1 to the end). The input features were the same as for the energy estimator.

5.2.3. Context matching

Bursts are subject to co-articulation like any other speech sounds. For example, a /k/ sound in the neighborhood of a rounded vowel will inherit some of the spectral characteristics of the vowel.
We use a simplified context matching approach in these experiments, where only four classes are considered:

- rounded vowel
- unrounded vowel
- consonant
- pause

Thus, a given candidate burst in the library, with a left- and right context B_{left} , B_{right}, can be placed in the context P_{left} , P_{right} it is considered to match if

$$\text{class}\{B_{left}\}=\text{class}\{P_{left}\} \;\&\; \text{class}\{B_{right}\}=\text{class}\{P_{right}\}.$$

5.2.4. Selecting the best candidate

We start with a list of candidates made up of all bursts that match the phonetic context. This means that 1) the phone itself should match, and 2) the immediate context should match as defined in 5.2.3. Then we consider the energy estimated by the regression tree. If it falls below a certain threshold, no burst will be placed in the specific location, otherwise we calculate an energy score for each candidate *k* according to

$$S_{energy}(k) = 1 - (1/3)\,|E_k - \hat{E}|$$

where Ek is the energy class value for candidate k stored in the library and $\hat{E}$ is the estimated energy predicted by the contextual features. In a similar way we calculate a score for how well the duration of the target segment matches:

$$S_{dur}(k) = 1 - |D_k - D|$$

where D_k is the duration of the segment from which the candidate was taken and D is the duration of the current segment in the synthesis specification.

The candidate with the highest combined score

$$S = S_{energy} + S_{dur}$$

is selected and placed at time

$$t = t_0 + \hat{O}\, D$$

where t_0 is the start of the current segment in the synthesis specification and $\hat{O}$ is the onset value predicted by the onset regression tree. Candidates that are too long to fit within the segment are discarded (i.e. a burst can be at most $(1\text{-}\hat{O})D$ seconds long without spilling into the next segment).

When the winning candidates have been selected for all the segments, the burst track is synthesized by placing the corresponding waveforms in the burst library at the specified points in time. Finally, the output speech waveform is produced by mixing together the harmonic track and the burst track.

6. Evaluation

In order to evaluate the proposed method, we set up a perceptual experiment to compare the synthesis results to that of a state of the art statistical parametric synthesizer.

6.1. Method

We selected a set of 50 consonant-rich sentences were selected from the set of Harvard sentences [17]. Each sentence was synthesized in two variants:

HTS: Using a standard voice built by the demo scripts for HTS ver. 2.3 (not using Straight)

HB: Using the harmonics + burst method proposed in this paper.

We used *Flite+hts_engine* [18] to synthesize the HTS voice and the harmonics part of the HB voice, and also to generate the full context label files for input to the burst synthesis in the HB voice.

An experiment was set up on the crowd sourcing platform *crowdflower.com* where people were invited to listen to the *HTS* and *HB* renditions of each sentence and answer two questions:

1. *Which one sounds more natural?*
2. *Which one is easier to understand?*

Each question had the alternatives "the first one", "the second one" and "I cannot hear any difference".

The order in which the two variants were played was varied between trials. A control stimuli was inserted in every block of 10 trials, where a natural sentence from the SLT Arctic database was to be compared to a *HTS* rendition of the same sentence.

Subjects were required to user headphones and to reside in English speaking countries (United States, United Kingdom, Australia, Canada) or top ranking countries in *EF English Proficiency Index*[1] (Sweden, Netherlands, Norway, Denmark, Finland). Every sentence pair was judged by 40 subjects, yielding a total of 2000 judgements, excluding the control stimuli. Subjects were paid 0.4 USD per block of 10 sentences. The control stimuli was used to filter out unreliable subjects: all users who responded that the Natural speech stimuli sounded less natural than the HTS rendition were discarded, leaving 1224 judgements.

6.2. Results

The distribution of responses can be seen in figure 2. Both for naturalness and intelligibility the proposed Harmonics-and-Bursts method (HB) scores higher than the comparison (HTS) on the average. The difference in the Naturalness score is substantial, while the difference in the Intelligibility score is marginal. A t-test reveals that the difference in Naturalness is significant ($p < 0.0002$) but the difference in intelligibility is not. It is however clear that it was a difficult task; both for Naturalness and Intelligibility the largest response category is "*I cannot hear any difference*".

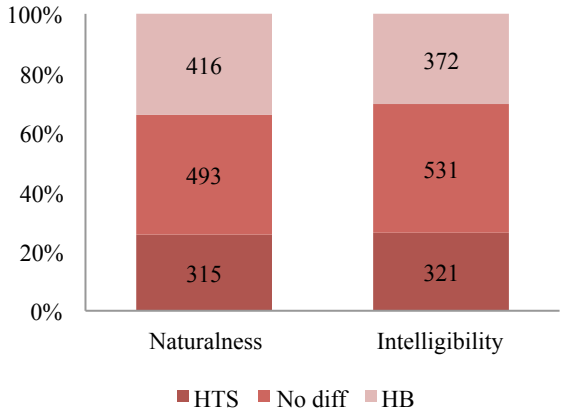

Figure 2: Results from the perceptual experiment where the proposed Harmonics-and-Bursts method HB is compared against HTS.

7. Conclusions

The idea of modeling different aspects of the speech signal along different time scales or using different methods makes sense conceptually, and has been proposed many times in the past. In [19] Sven Öhman suggested that consonant articulations may be viewed as rapid transients superimposed on a more slowly varying vowel track. This paper treats harmonic and burst signals in a similar way, and demonstrates that this indeed can be used to increase the naturalness of synthetic speech. One aspect of the work that requires further investigation is the effect on the filtering on the modeling of the harmonic signal. It is our subjective opinion that certain aspects of F0 extraction and spectral modeling improves when the noise bursts are removed before modeling, but this is a matter that requires further investigation. It is important to stress that the system described contains no manual steps and as such could be fully automated.

[1] http://www.ef.se/epi/

8. Acknowledgements

This work is backed by *ICT The Next Generation* and the The *Swedish Post and Telecom Authority (PTS)* via the *Wikispeech* project.

9. References

[1] Fant, G. (1971). *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations* (Vol. 2). Walter de Gruyter.

[2] Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *the Journal of the Acoustical Society of America*, *67*(3), 971-995.

[3] Carlson, R., Granström, B., & Karlsson, I. (1991). Experiments with voice modelling in speech synthesis. *Speech communication*, *10*(5), 481-489.

[4] Pearson, S., Holm, F., & Hata, K. (1997). Combining concatenation and formant synthesis for improved intelligibility and naturalness in text-to-speech systems.*International Journal of Speech Technology*, *1*(2), 103-107.

[5] Hirai, T., Yamagishi, J., & Tenpaku, S. (2007). Utilization of an HMM-based feature generation module in 5 ms segment concatenative speech synthesis. In*SSW* (pp. 81-84).

[6] Rouibia, S., & Rosec, O. (2005). Unit selection for speech synthesis based on a new acoustic target cost. In *Ninth European Conference on Speech Communication and Technology*.

[7] Ling, Z. H., & Wang, R. H. (2008, March). Minimum unit selection error training for HMM-based unit selection speech synthesis system. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*(pp. 3949-3952). IEEE.

[20]

[8] Gonzalvo, X., Gutkin, A., Socoró, J. C., Sanz, I. I., & Taylor, P. (2009). Local minimum generation error criterion for hybrid HMM speech synthesis. In *INTERSPEECH* (pp. 416-419).

[9] Tiomkin, S., Malah, D., Shechtman, S., & Kons, Z. (2011). A hybrid text-to-speech system that combines concatenative and statistical synthesis units.*Audio, Speech, and Language Processing, IEEE Transactions on*, *19*(5), 1278-1288.

[10] Sorin, A., Shechtman, S., & Pollet, V. (2011). Uniform Speech Parameterization for Multi-Form Segment Synthesis. In *INTERSPEECH* (pp. 337-340).

[11] Raitio, T., Suni, A., Pulakka, H., Vainio, M., & Alku, P. (2011, May). Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on* (pp. 4564-4567). IEEE.

[12] Liutkus, A., Rafii, Z., Pardo, B., Fitzgerald, D., & Daudet, L. (2014, May). Kernel Spectrogram models for source separation. In *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on*(pp. 6-10). IEEE.

[13] https://github.com/google/REAPER, retrieved 2015-02-17.

[14] Kominek, J., & Black, A. W. (2004). The CMU Arctic speech databases. In *Fifth ISCA Workshop on Speech Synthesis*.

[15] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., & Tokuda, K. (2007, August). The HMM-based speech synthesis system (HTS) version 2.0. In *SSW* (pp. 294-299).

[16] Taylor, P., Caley, R., Black, A. W., & King, S. (1999). Edinburgh speech tools library. *System Documentation Edition*, *1*, 1994-1999.

[17] Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., & Weinstock, M. (1969). IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio Electroacoust*, *17*(3), 225-246.

[18] http://hts-engine.sourceforge.net

[19] Öhman, S. E. (1967). Numerical model of coarticulation. *The Journal of the Acoustical Society of America*, *41*(2), 310-320.