Towards Personalized Synthesized Voices for Individuals with Vocal Disabilities: Voice Banking and Reconstruction

Christophe Veaux, Junichi Yamagishi, Simon King

Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
{cveaux, jyamagis}@inf.ed.ac.uk, Simon.King@ed.ac.uk

Abstract

When individuals lose the ability to produce their own speech, due to degenerative diseases such as motor neurone disease (MND) or Parkinson's, they lose not only a functional means of communication but also a display of their individual and group identity. In order to build personalized synthetic voices, attempts have been made to capture the voice before it is lost, using a process known as voice banking. But, for some patients, the speech deterioration frequently coincides or quickly follows diagnosis. Using HMM-based speech synthesis, it is now possible to build personalized synthetic voices with minimal data recordings and even disordered speech. The power of this approach is that it is possible to use the patient's recordings to adapt existing voice models pre-trained on many speakers. When the speech has begun to deteriorate, the adapted voice model can be further modified in order to compensate for the disordered characteristics found in the patient's speech. The University of Edinburgh has initiated a project for voice banking and reconstruction based on this speech synthesis technology. At the current stage of the project, more than fifteen patients with MND have already been recorded and five of them have been delivered a reconstructed voice. In this paper, we present an overview of the project as well as subjective assessments of the reconstructed voices and feedback from patients and their families.

Index Terms: HTS, Speech Synthesis, Voice Banking, Voice Reconstruction, Voice Output Communication Aids, MND.

1. Introduction

Degenerative speech disorders have a variety of causes that include Multiple Sclerosis, Parkinson's, and Motor Neurone Disease (MND) also known in the USA as Amyotrophic Lateral Sclerosis (ALS). MND primarily affects the motor neurones in the brain and spinal cord. This causes a worsening muscle weakness that leads to a loss of mobility and difficulties with swallowing, breathing and speech production. Initial symptoms may be limited to a reduction in speaking rate, an increase of the voice's hoarseness, or an imprecise articulation. However, at some point in the disease progression, 80 to 95% of patients are unable to meet their daily communication needs using their speech; and most are unable to speak by the time of their death [1]. As speech becomes difficult to understand, these individuals may use a voice output communication aid (VOCA). These devices consist of a text entry interface such as a keyboard, a touch screen or an eye-tracker, and a text-to-speech synthesizer that generates the corresponding speech. However, when individuals lose the ability to produce their own speech, they lose not only a functional means of communication but also a display of their individual and social identity through their vocal characteristics.

Current VOCAs are not ideal as they are often restricted to a limited set of impersonal voices that are not matched to the age or accent of each individual. Feedback from patients, careers and patient societies has indicated that there is a great unmet need for personalized VOCAs as the provision of personalized voice is associated with greater dignity and improved self-identity for the individual and their family [2].

In order to build personalized VOCAs, several attempts have been made to capture the voice before it is lost, using a process known as voice banking. One example of this approach is ModelTalker [3], a free voice building service that can be used from any home computer in order to build a synthetic voice based on diphone concatenation, a technology developed in the 1980s. The user of this service has to record around 1800 utterances in order to fully cover the set of diphones and the naturalness of the synthetic speech is rather low. Cereproc [4] has provided a voice building service for individuals, at a relatively high cost, which uses unit selection synthesis, and is able to generate synthetic speech of increased naturalness. Wants Inc. in Japan also provides a commercial voice building service for individuals called "Polluxstar". This is based on a hybrid speech synthesis system [5] using both unit selection and statistical parametric speech synthesis [6] to achieve a natural speech quality. However, all these speech synthesis techniques require a large amount of recorded speech in order to build a good quality voice. Moreover the recorded speech data must be as intelligible as possible, since the data recorded is either used directly or partly as the voice output. This requirement makes such techniques more problematic for those patients whose voices have started to deteriorate. Therefore, there is a strong motivation to reduce the complexity and to increase the flexibility of the voice building process so that patients can have their own synthetic voices build from limited recordings and even deteriorating speech.

Recently, a new voice building process using the hidden Markov model (HMM)-based speech synthesis technique has been investigated to create personalized VOCAs [7-8]. This approach has been shown to produce high quality output and offers two major advantages over existing methods for voice banking and voice building. First, it is possible to use existing speaker-independent voice models pre-trained over a number of speakers and to adapt them towards a target speaker. This process known as speaker adaptation [9] requires only a very small amount of speech data. The second advantage of this approach is that we can control and modify various components of the adapted voice model in order to compensate for the disorders found in the patient's speech. We call this process "voice reconstruction". Based on this new approach, the University of Edinburgh, the Euan MacDonald Center for MND and the Anne Rowling Regenerative Neurology Clinic have started a collaborative

SLPAT 2013, 4th Workshop on Speech and Language Processing for Assistive Technologies, pages 107–111,
Grenoble, France, 21–22 August, 2013.

project for voice banking and voice reconstruction [10-11]. At the current stage of the project, more than 15 patients with MND have already been recorded and 5 of them have been delivered a reconstructed voice. We present here the technical concepts behind this project as well as a subjective assessment of the reconstructed voices.

2. HMM-Based Speech Synthesis

Our voice building process is based on the state-of-the-art HMM-based speech synthesizer, known as HTS [6]. As opposed to diphone or unit-selection synthesis, the HMM-based speech synthesizer does not use the recorded speech data directly as the voice output. Instead it is based on a vocoder model of the speech and the acoustic parameters required to drive this vocoder are represented by a set of statistical models. The vocoder used in HTS is STRAIGHT and the statistical models are context-dependent hidden semi-Markov models (HSMMs), which are HMMs with explicit state duration distributions. The state output distributions of the HSMMs represent three separate streams of acoustic parameters that correspond respectively to the fundamental frequency (logF0), the band aperiodicities and the mel-cepstrum, including their dynamics. For each stream, additional information is added to further describe the temporal trajectories of the acoustic parameters, such as their global variances over the learning data. Finally, separate decision trees are used to cluster the state durations probabilities and the state output probabilities using symbolic context information at the phoneme, syllable, word, and utterance level. In order to synthesize a sentence, a linguistic analyser is used to convert the sequence of words into a sequence of symbolic contexts and the trained HSMMs are invoked for each context. A parameter-generation algorithm is then used to estimate the most likely trajectory of each acoustic parameter given the sequence of models. Finally the speech is generated by the STRAIGHT vocoder driven by the estimated acoustic parameters.

3. Speaker Adaptation

One advantage of the HMM-based speech synthesis for voice building is that the statistical models can be estimated from a very limited amount of speech data thanks to speaker adaptation. This method [9] starts with a speaker-independent model, or "**average voice model**", learned over multiple speakers and uses model adaptation techniques drawn from speech recognition such as maximum likelihood linear regression (MLLR), to adapt the speaker independent model to a new speaker. It has been shown that using 100 sentences or approximately 6-7 minutes of speech data is sufficient to generate a **speaker-adapted voice** that sounds similar to the target speech [7]. This provides a much more practical way to build a personalized voices for patients. For instance, it is now possible to construct a synthetic voice for a patient prior to a laryngectomy operation, by quickly recording samples of their speech [8]. A similar approach can also be used for patients with degenerative diseases before the diseases affect their speech. The speaker adaptation process is most successful when the average voice model is already close to the voice characteristics of the target speaker. Therefore, one goal of the voice-bank project is to record a large catalogue of healthy voices from which we can derive a set of average voice models corresponding to different age, gender and regional accents combinations. This will be presented in Section 5.

4. Voice Reconstruction

Some individuals with neurodegenerative disease may already have speech symptoms at the time of the recording. In that case, the speaker adaptation process will also replicate these symptoms in the speaker-adapted voice. Therefore we need to remove speech disorders from the synthetic voice, so that it sounds more natural and more intelligible. However since the HTS is based on a vocoder model of the speech, we can now exploit the acoustic models learned during the training and the adaptation processes in order to control and modify various speech features. This is the second major advantage of using HMM-based speech synthesis. In particular, HTS has statistically independent models for duration, log-F0, band aperiodicity and mel-cepstrum. This allows the substitution of some models in the patient's speaker-adapted voice by that of a well-matched healthy voice or an average of multiple healthy voices, as illustrated in Figure 1. Although disordered speech perceptually deviates considerably from normal speech in many ways, it is known that its articulatory errors are consistent [12] and hence relatively predictable [13]. Therefore we can pre-define a substitution strategy for a given condition, to some extent.

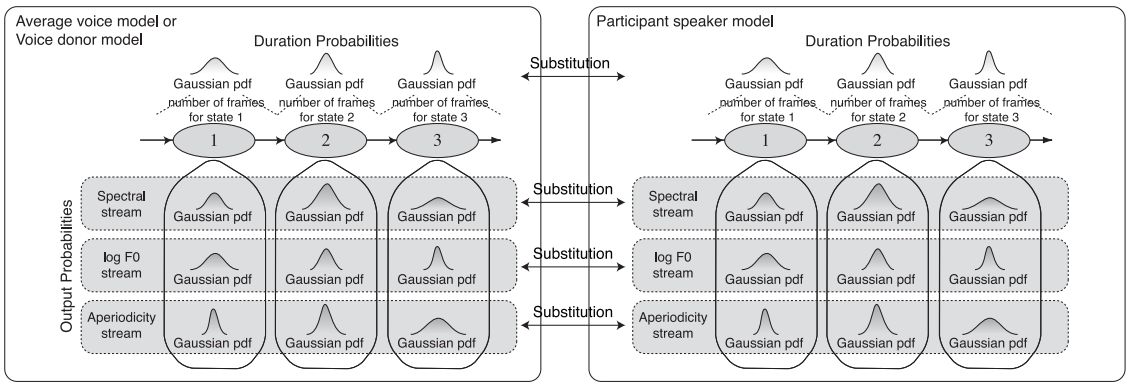

Figure 1: *The structure of the acoustic models in HTS means that there can be a substitution of state output or state duration models between an healthy voice model and the patient voice model in order to compensate for any deterioration in the patient's speech.*

For example, patients with MND often have a disordered speaking rate, contributing to a loss of the speech intelligibility. The substitution of the state duration models enables the timing disruptions to be regulated at the phoneme, word, and utterance levels. Furthermore, MND speakers often have breathy or hoarse speech, in which excessive breath through the glottis produces unwanted turbulent noise. In such cases, we can substitute the band aperiodicity models to produce a less breathy or hoarse output. In the following part of this section, we present different levels of model substitution. All these levels are combined in the final voice reconstruction process.

4.1. Baseline model substitution

In a first approach [7], the following models and information are substituted:

- Duration and aperiodicity models
- Global variances of log-F0, aperiodicity and mel-cepstrum

These parameters are the less correlated with the speaker identity and their substitution can fix some disorders such as slow speaking rate and excessive hoarseness. However, this substitution strategy cannot correct articulation disorders.

4.2. Component-wise model substitution

This is an extension of the baseline model substitution. Since the state output distributions have diagonal covariance matrix, we can substitute a component independently from the others. This component-wise substitution strategy allows to substitute the parts of the mel-cepstrum and log-F0 streams that are the less correlated with the speaker identity. In this way, we can further reduce some disorders without altering the voice identity. In particular, we substitute the mean and variance for the following components:

- 1st coefficient of the mel-cepstrum (energy)
- High-order coefficients of the mel-cepstrum
- Dynamics coefficients of the mel-cepstrum and log-F0
- Voiced/Unvoiced weights

The substitution of the high order static coefficients and the dynamics coefficients of the mel-cepstrum will help to reduce the articulation disorders without altering the timbre. In our implementation, we replace all static coefficients of order N>40. The substitution of the dynamics coefficients of the log-F0 will help to regulate the prosodic disorders such as monotonic F0. Finally the replacement of the voiced/unvoiced weights will fix the breathiness disorders. The duration models, aperiodicity models, and global variances are also substituted as in the baseline strategy. We will refer to this method as the **component-wise strategy**.

4.3. Context-dependent model substitution

In the two previous strategies, the model substitutions are independent of the context. However, in HTS, the acoustic models are clustered after their contexts by separate decisions trees. We can use this contextual information to further refine the model substitution. For example, some MND patients cannot pronounce correctly the plosives, the approximants and the diphthongs. In these contexts, it is preferable to substitute all the mel-cepstrum coefficients in order to enhance the intelligibility of the speech. Therefore, we have defined a **context-dependent strategy**, in which the mel-cepstrum models are entirely substituted for some specific contexts. Since these contexts may vary from one patient to the other, we have designed a screening procedure in which the patients have to read out a set of 50 sentences covering most of the phonetic contexts. Their speech is then assessed by a speech therapist in order to define the contexts for which the models are to be substituted. Finally, the context-dependent and the component-wise model substitutions are combined in order to get the final version of the repaired voice. Ideally the voice donors used for the voice reconstruction should share the gender, age range and regional accent of the patient since these factors are likely to contribute to the characteristics of the voice. This is why we need to record a large number of healthy voice donors with a variety of age and regional accents, as presented in the next section.

5. Database of Voice Donors

One of the key elements of the voice-banking project is the creation of a catalogue of healthy voices with a wide variety of accents and voice identities. This voice catalogue is used to create the average voice models for the speaker adaptation and to select the voice donors for the voice reconstruction. So far we have recorded about 500 healthy voice donors with various accents (Scottish, Irish, Other UK). This database is already the largest UK speech research database. An illustration of the geographical distribution of the speakers' birthplaces is shown on Figure 2. Each speaker has been recorded in a semi-anechoic chamber for about one hour using at each time a different script in order to get the best phonetic coverage on average. The database of healthy voices is first used to create the average voice models used for speaker adaptation. Ideally, the average voice model should be close to the vocal identity of the patient and it has been shown that gender and regional accent are the most influent factors in speaker similarity perception [14]. Therefore, the speakers are clustered according to their gender and their regional accent in order to train specific average voice models. A minimum of 10 speakers is required in order to get robust average voice models.

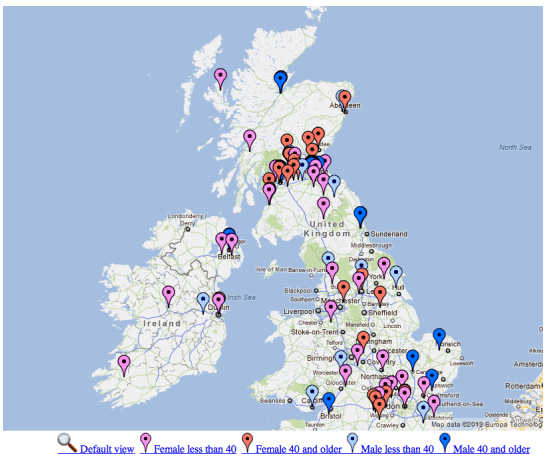

Figure 2: UK-wide speech database.

The healthy voice database is also used to select the voice donors for the model substitution process described in section 4. The voice donors are chosen among the speakers used to build the average voice model matched to the patient's gender and accent.

We first build a speaker-adapted voice for each of these speakers using the same average voice model. The acoustic models used in HTS represent each stream of parameters separately. Therefore, a set of acoustic distances between speaker-adapted voices can be defined for each of these streams (duration, log-F0, band aperiodicity, mel-cepstrum). These distances are defined as the average Karhunen-Loeve (KL) distances [15] between the acoustics models associated to the same stream of parameters. Finally, a voice donor is selected for each stream separately, as the one that minimizes the average acoustic distance for this stream.

6. Clinical Trial

As part of the voice-banking project, we are conducting a clinical trial in order to assess and further refine the voice building process for patients with degenerative speech disorders. So far, more than 15 patients with MND have already been recorded and 5 of them have been delivered a reconstructed voice. We present in the following sections a subjective assessment of the voice repair as well as the feedbacks from patients and their families.

4.3. Subjective evaluation of the voice repair

The substitution strategy presented in Section 4 was evaluated for the case of a MND patient. This patient was a 45 years old Scottish male that we recorded twice. A first recording of one hour (500 sentences) has been made just after diagnosis when he was at the very onset of the disease. At that time, his voice did not show any disorders and could still be considered as "healthy". A second recording of 15 minutes (50 sentences) has been made 10 months later. He has then acquired some speech disorders typically associated with MND, such as excessive hoarseness and breathiness, disruption of speech fluency, reduced articulation and monotonic prosody. The synthetic voices used in this experiment are shown in Table 1. The same male-Scottish average voice model, denoted as AV, was used to create all the synthetic voices. This average voice was trained on 17 male Scottish speakers using 400 sentences each giving a total of 6800 sentences. The synthetic voice created from the first recording of the patient ("healthy" speech) was used as the reference voice for the subjective evaluations. This reference voice is referred to as HC. This choice of a synthetic voice as reference instead of the natural recordings was done to avoid any bias due to the loss of quality inherent to the synthesis. The reconstructed voice IR was obtained by applying the combination of the **component-wise** and **context-dependent** substitution strategies to the speaker-adapted voice IC build from the second recording of the patient ("impaired" speech).

Voice	*Description*
AV	Average voice used for speaker adaptation
HC	Speaker adapted voice of the "**healthy**" speech
IC	Speaker adapted voice of the "**impaired**" speech
IR	Reconstructed voice using the **component-wise** and **context-dependent** model substitutions

Table 1: Voices compared in the evaluation tests

In order to evaluate the effectiveness of the voice reconstruction, two subjective tests were conducted. The first one assesses the intelligibility of the synthesized voice and the second, the speaker similarity. The same 40 semantically unpredictable sentences [16] were synthesized for each of the 3 voices created from the patient's recordings (see Table 1). The resulting synthesized samples were divided into 4 groups such that each voice is represented by 10 samples in a group. A total of 40 native English participants were asked to transcribe the synthesized samples, with 10 participants for each group. Within each group, the samples were presented in random order for each participant. The participants performed the test with headphones. The transcriptions were evaluated by measuring the word error rate (WER).

Voice	*Mean WER (%)*	*std*
HC	26	12
IC	53	18
IR	36	16

Table 2: Word Error Rate (mean, standard deviation)

The same test sentence "People look, but no one ever finds it." was synthesized for each of the 4 voices in Table 1. Participants were asked to listen alternatively to the reference voice HC and to the same sentence synthesized with the reconstructed voice IR and the average voice model AV. The presentation order of the voices being tested was randomized. Participants should rate the similarity between the tested voice and the reference HC on a 5-point scale (1: Very dissimilar, 2: Dissimilar, 3: Quite Similar, 4: Very similar; and 5: Identical). However, the participants were not given further instruction in order to avoid biasing towards rating any specific form of similarity. A total of 40 native English speakers performed the test using headphones.

Voice	*Mean Opinion Score*	*std*
AV	2.05	1.05
IC	2.61	1.21
IR	3.09	1.34

Table 3: Similarity to the reference voice HC on a MOS-scale (mean, standard deviation)

The resulting average WERs for the intelligibility test are shown in Table 2. We are not interested here in the absolute values of the WER but in their relative values compared to the reference voice HC. As expected, the synthetic voice IC created from the "impaired" speech has a high WER, which means that the articulation disorders from the patient's speech have degraded the intelligibility. The important result here is that the model substitution improves the speech intelligibility of the reconstructed voice IR. The results of the similarity test are shown in Table 3. A first interesting result is that the voice clone IC created by speaker adaptation from the "impaired" speech is more similar to the healthy clone HC that the average voice AV. In the case of this patient, this validates an implicit assumption of the voice reconstruction process: some valuable information about the original vocal identity should remain in the impaired speech. The other important result is the improvement of the average similarity scores when the model substitution strategies are applied. Between IR and AV, there is a mean improvement of 1 MOS (with a p-value << 1.e-5) and more surprisingly, there is also a significant improvement of 0.5 MOS (p-value << 1.e-3) between IC and IR. One explanation of this last result could be

that the similarity of vocal identity is better perceived once the disorders have been regulated.

4.3. Feedback from patients and families

The results presented in the previous section are relative to the only patient whose 'healthy' voice was available to establish a reference. However, it remains to be demonstrated that similar results could be achieved with different patients. It is also important to assess the usability of the reconstructed voice in real conditions of use. Therefore, we have conducted an experimental trial with 5 patients whose voices have been reconstructed and made available through an on-line server. These patients can use their reconstructed voices from any computer, tablet or mobile phone as long as an Internet connection is available. A simple web interface allows them to enter a text and a synthesis request is sent to a remote server. Once the synthesis is done on the server, the synthesized speech is sent to the device and played through its loudspeakers. The patients and their families were asked to give their feedback on the quality of the reconstructed voice after a few weeks of use. In particular, they were asked to assess the intelligibility of the voice and its similarity to the user's voice before the start of the disease. We get 15 feedback in total corresponding to the 5 patients, their husbands/wives or their parents. The table 4 shows the mean opinion scores on a 5-point scale (1 being the worst and 5 the best). These results are consistent with the subjective test presented in the previous section. It shows that the voice reconstruction process manages to remove most of the speech artifacts while retaining some of the voice characteristics of the patient. Most importantly, all the patients said they would rather choose their reconstructed voices over any commercially available synthetized voice.

Question	*Mean Opinion Score*	*std*
Similarity	3.5	0.7
Intelligibility	4	1.1

Table 4: Feedback from patients and families (mean, standard deviation)

7. Conclusions

For VOCA users, speech synthesis is not an optional extra for reading out text, but a critical function for social communication and identity display. Therefore, there is a great need for personalized VOCAs as the provision of personalized voice is associated with greater dignity and improved self-identity for the individual and their family. In order to build personalized synthetic voices, attempts have been made to capture the voice before it is lost, but for some patients, the speech deterioration frequently coincides or quickly follows diagnosis. In such cases, HMM-based speech synthesis has two clear advantages: speaker adaptation and improved control. Speaker adaptation allows the creation of a synthetic voice with a limited amount of data. Then the structure of the acoustic models can be modified to repair the synthetic speech. In this paper, we have presented the results of an on-going clinical trial based on this new approach. The subjective evaluations and the feedback from the patients show that it is possible to build a synthesized voice that retains the vocal identity of the patient while removing most of the speech disorders. Although these results are presented for MND patients, the principle of the voice building and reconstruction process could be easily generalized to any other degenerative or acquired speech disorder.

8. References

[1] Doyle, M. and Phillips, B. (2001), "Trends in augmentative and alternative communication use by individuals with amyotrophic lateral sclerosis," *Augmentative and Alternative Communication* 17 (3): pp.167–178.

[2] Murphy, J. (2004), "I prefer this close': Perceptions of AAC by people with motor neurone disease and their communication partners. *Augmentative and Alternative Communication*, 20, 259-271.

[3] Yarrington, D., Pennington, C., Gray, J., & Bunnell, H. T. (2005), "A system for creating personalized synthetic voices," *Proc. of ASSETS*.

[4] http://www.cereproc.com/

[5] Kawai, H., Toda, T., Yamagishi, J., Hirai, T., Ni, J., Nishizawa, N., Tsuzaki, M., and Tokuda, K. (2006) "XIMERA: a concatenative speech synthesis system with large-scale corpora," IEICE Trans. Information and Systems, J89-D-II (12), pp.2688–2698.

[6] Zen, H., Tokuda, K., & Black, A. (2009) "Statistical parametric speech synthesis, Speech Communication," 51, pp.1039-1064.

[7] Creer, S., Green, P., Cunningham, S., & Yamagishi, J. (2010) "Building personalized synthesized voices for individuals with dysarthia using the HTS toolkit," IGI Global Press, Jan. 2010.

[8] Khan, Z. A., Green P., Creer, S., & Cunningham, S. (2011) "Reconstructing the Voice of an Individual Following Laryngectomy," Augmentative and Alternative Communication.

[9] Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K. & Isogai, J. 2009. Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans. on ASL, 17, 66-83.

[10] Veaux, C., Yamagishi, J., King, S. (2011) "Voice Banking and Voice Reconstruction for MND patients," *Proceedings of ASSETS*.

[11] Veaux, C., Yamagishi, J., King, S. (2012) "Using HMM-based Speech Synthesis to Reconstruct the Voice of Individuals with Degenerative Speech Disorders," *Interspeech*, Portland, USA.

[12] Yorkston, K. M., Beukelman, D. R. and Bell, K. R. (1998) "Clinical management of dysarthric speakers," College-Hill Press.

[13] Mengistu, K.T. and Rudzicz, F., (2011) "Adapting acoustic and lexical models to dysarthric speech," Proc. ICASSP 2011.

[14] Dall, R., Veaux, C., Yamagishi, J. & King, S. (2012) "Analysis of speaker clustering strategies for HMM-based speech synthesis," *Proc. Interspeech*, Portland, USA.

[15] Trung Hieu Nguyen, Haizhou Li, and Eng Siong Chng. (2009) "Cluster criterion functions in spectral subspace and their application in speaker clustering," In *Proceedings of ICASSP*.

[16] Benoît C., Grice M., & Hazan, V. (1996) "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences," Speech Communication.