Manipulating tracheoesophageal speech

van Son, R.J.J.H.; Jacobi, I.; Hilgers, F.

Publication date
2010

Document Version
Author accepted manuscript

Published in
Interspeech

Link to publication

Citation for published version (APA):
Manipulating Treacheoesophageal Speech

R.J.J.H. van Son, Irene Jacobi, and Frans Hilgers

Netherlands Cancer Institute/ACLC, University of Amsterdam, The Netherlands

R.J.J.H.vanSon@gmail.com

Abstract

Speech therapy aiming at improving voice quality and speech intelligibility is often hampered by the lack of knowledge of the underlying deficits. One way to help speech therapists treating patients would be to supply synthetic benchmarks for pathological speech. These can be used to train therapists and evaluate and interpret automatic speech recognizers used for diagnosing pathological speech. Moreover, synthetic pathological speech can also be used to make expected therapy aims audible before treatment. In a listening experiment testing perceived intelligibility, three types of manipulations of tracheoesophageal speech were evaluated by experienced speech therapists. It was found that modeling the intensity contour of the voice source signal improved speech quality over plain analysis-synthesis. Replacing the voicing source with fully synthetic source periods decreased the perceived intelligibility markedly. Making the source fully periodic with a regular pitch had no effect on perceived intelligibility. Low quality speech benefitted more from manipulations, or deteriorated less, than high quality speech.

Index Terms: Treacheoesophageal speech, pathological speech synthesis

1. Introduction

Pathologic speech arising from oncologic treatment has a significant negative impact on the quality of life (QoL) of patients. Therefore, clinical interventions and surgical treatments aim at the preservation of speech quality as much as possible. And if preservation is no longer possible, like in advanced larynx cancer, and a complete laryngectomy is unavoidable, efforts are made to improve the substitute voice.

Presently, the prospects for the development of an adequate substitute voice due to the use of prosthetic devices are good, e.g., [1, 2, 3]. Subsequent speech therapy will then aim at further improving voice quality and speech intelligibility. Studies have shown that improvements of speech quality and intelligibility can indeed dramatically improve the QoL of patients [2]. To support and evaluate therapies, efforts have been recently made to introduce objective methods and automatic evaluations of the intelligibility and quality of alaryngeal speech [4, 5].

The interpretation of automated, e.g., ASR based, evaluation results requires knowledge about the relation between speech quality, medical history, and physiological characteristics of the patient. Such knowledge could also be used to develop models of a patient’s speech during therapy. Eventually enabling to predictively synthesize post intervention speech and an evaluation of the improvement of individual speech qualities, e.g., intelligibility, even before intervention starts.

For the clinical practice, the re-creation of specific speech pathologies, or even the patient voice is relevant, c.f., [6]. Using an analysis-by-synthesis approach, possible causes of patient specific problems might be identified [6, 7, 8, 9, 10]. If the resulting synthetic voice is acceptable, the effects of specific therapies might then be tested on manipulated synthesis parameters.

However, the methods described in the literature are generally targeted at full blown (formant) synthesis of the speech with a replaced voice source. Although this would be the ideal procedure, there are currently too many problems with analysis-synthesis to make this method practical for daily use, especially for TE speech. For one thing, current understanding of the TE source is not at the same level of detail as the (ab-) normal glottal source. Problems of TE speakers are often concentrated in the area of voicing distinctions and voice stability [11, 12, 3]. Therefore, it might be preferable to only manipulate specific aspects of the synthetic source in an analysis-synthesis procedure of individual patients to evaluate the effects of individual factors, like voicing distinctions and pitch stability.

The current study manipulates TE speech of individual patients to evaluate how voice source parameters affect perceived speech intelligibility. Three problematic aspects of the TE voice source are manipulated: Amplitude stability, pitch stability, and source spectrum (by way of pitch period shape).

1.1. Creating analysis-synthesis speech

Recreating new [7, 8, 9] or enhanced [6, 10] pathological speech is an active area of research. Manipulations are generally based on a simple Source-Filter model of speech [13]. The preferred method is to create a synthetic voice source signal, either de novo or from actual speech and to filter it with a model of the vocal tract filter. The model of the vocal tract filter too can be entirely synthetic or derived from actual speech. The full cycle is to analyze speech and extract a filter and source. Then the source or filter are manipulated followed by filtering the new source with the new filter, to (re-)synthesize the speech.

The current study uses analysis-synthesis based on LPC (formant) filters and an inverse filtered LPC-source signal. The ultimate goal is to be able to understand and manipulate the LPC-source and LPC-filter for each individual patient. That is, to replace the LPC-source, or filter, with a synthetic version that can be tuned to the outcome of different therapies and treatment options to predict speech quality.

LPC analysis models speech with an independent source spectrum (by way of pitch period shape).
with the original recordings. One particular problem is that LPC analysis is based on the assumption that the source signal is spectrally “white”, i.e., flat. The real human voicing source has a “pink” spectrum, i.e., it falls off with approximately 3dB per octave. This cannot be corrected perfectly. As a result, LPC synthesis tends to sound rather “dull” (de-emphasized) or overly sharp and noisy (emphasized). The high frequency noise of the emphasized correction was considered more harmful. So the de-emphasized synthesis was chosen for the experiments. All stimuli were created using this analysis-synthesis route to allow a fair comparison.

The source filter model, and therefore, inverse filtering, only holds for speech produced with a source at the (neo-) glottis. If a source signal is introduced elsewhere, e.g., in fricatives and plosives, this simple model will break down. It is crucial to detect the unvoiced parts as the LPC synthesis of unvoiced speech has unwanted effects on perceived qualities and intelligibility. In the current study, unvoiced parts of the original speech are transmitted unaltered as in [6, 7, 8].

An important complication when using the LPC-source to manipulate alaryngial speech is the difficulty in determining what parts of the speech are produced voiced and what unvoiced [6, 10]. The voiced/unvoiced decision is not always trivial in normal speakers, it can become extremely difficult and unreliable in TE speakers. Therefore, all voiced/unvoiced distinction were made by hand.

2. Materials and Methods

2.1. Speech recordings

16 male alaryngeal TE speakers, age 46-82 (median age 58), read aloud a short magazine story on two different occasions. These readings were recorded as part of their speech therapy sessions. All speakers have given informed consent which make these recordings available for research within the institute. In total 31 recordings of the short sentence ook het WEER heeft aan deze tocht meegewerkt (English: The weather has also contributed to this trip) from the end of the story were extracted. Two speakers failed to include this sentence in one of the sessions. One speaker read it twice during a session. One of these double readings was used as a practice item. In total 30 recordings from 16 speakers were used to generate the stimuli for this experiment.

2.2. Stimulus generation

All analysis, synthesis, and manipulations are done using Praat 2.2.4. Stimuli with fully synthetic source (NS)

Complete replacement of the source is based on the pitch pulse markers from the original sound. Each pair of pitch pulse marks are replaced with a standard source (differentiated) pitch period (Open phase 0.7, Collision phase 0.03, Power1 3.0, Power2 4.0 [14]). This synthetic source is multiplied with the original Intensity contour and de-emphasized (from 50Hz) before synthesis.

2.3. Subjects and listening experiment

Six subjects participated. All were native speakers of Dutch and had ample experience with TE speech. Four are speech therapists who work, or have worked, at the Dutch Cancer Institute.
performed the on-line experiment in a single half hour session.

3. Results

All statistics are done with R [17]. The overall distribution of intelligibility scores is presented in figure 1. The responses to the baseline AS stimuli are spread over all 7 levels indicating a highly variable baseline speech quality. Subjects were able to consistently rate stimuli (p < 0.001 for each of AS, EI, EP, and NS; χ²=29, χ² > 99, Friedman rank sum test).

Responses to AS stimuli were taken as the baseline quality. This baseline value was subtracted from responses of the same subject to other recordings of this speech recording. That is, if the response of a subject to the AS rendering of a recording was 4, and her response to the EP rendering of the same recording was 2, then the difference for EP would be -2 (= 4 - 2).

The average differences for EI, EP, and NS are presented in figure 2. The EI stimuli were, on average, perceived to be more intelligible, and the NS stimuli less intelligible than the corresponding baseline AS stimuli. Both differences were statistically significant (p < 0.001, Wilcoxon Matched Pairs Signed Ranks test). There was no difference for the EP stimuli.

It has been noted before that high quality pathological speech benefits less, or deteriorates more, than poor quality speech from manipulations intended to improve the speech [6]. This was tested by calculating the correlation between the relative intelligibility scores and the absolute baseline scores. Different subjects will have different response biases which will affect the overall correlation. Therefore, for each subject, all scores were recalculated to Z-scores with a standard normal distribution. The correlations between absolute score for the AS baseline stimuli and the relative scores for the other, manipulated, stimuli were calculated on these standardized scores. These correlations are presented in figure 3.

There are obvious scale boundary effects on the correlations. The regression line under H₀, i.e., differences are uncorrelated to the AS baseline, was determined with a Monte-Carlo simulation. AS scores were combined with random uniform differences between [-2, 2] (~95% of observed differences), then “rounded” to the scale boundaries. Regression lines were calculated for 50,000 runs. The mean values of the simulation are plotted as H₀ in figure 3. The two-sided p values of the observed correlations are determined as percentiles of the Monte-Carlo simulation and were all significant (p < 0.001, figure 3). It is obvious that the impact of the manipulations became more negative with higher baseline AS quality, supporting [6].

Figure 1: Distribution of intelligibility scores for the baseline analysis-synthesis (AS) and manipulated (EI, EP, NS) stimuli. p < 0.001 Friedman rank sum test for each stimulus type. N: number of total responses

Figure 2: Mean difference in perceived intelligibility scores with respect to plain AS baseline stimuli (figure 1). Higher is better quality. EI: Equalized Intensity, EP: Equalized Pitch, NS: New Source signal. *: p < 0.001, ns: p ≥ 0.001. N: number of responses per stimulus type.

Figure 3: Correlation between difference scores (figure 2) and control AS scores (figure 1). Correlations are calculated on the standardized judgements (Z-scores). H₀: Monte-Carlo simulation of uncorrelated responses. See figure 2 and text.
seems to be easier to improve low quality speech than high quality speech. This holds for manipulations that improved speech intelligibility, EI, as well as those that made it worse, NS, or had no effect, EP. The effect was stronger for NS than EI stimuli (p < 0.005, two sided Fisher-Z transform).

4. Discussion and conclusions

TE speakers often have specific voice problems that impair their speech. The question then is what therapy would improve their speech most. It would be advantageous if treatment options could be evaluated in advance on their ability to reduce or even prevent these problems. In these cases, it would be useful if features of the speech could be selectively manipulated to evaluate how changes would affect speech quality.

The current paradigm for manipulating pathological speech in general, and TE speech in particular, is to model the vocal tract or voicing source from first principles, i.e., ab initio, [6, 7, 8, 9, 10]. Successful synthesis from first principles is, indeed, the method that would improve our understanding most. However, it is still very difficult, if not impossible, to recreate the speech of individual patients with a quality that would be useful in clinical practice.

The aim of the current study is to evaluate an alternative course of action. Instead of synthesizing the voicing source from first principles, only specific aspects of the existing recordings are manipulated. This way, the features that might be relevant to the problem at hand can be evaluated within the context of what could actually be addressed in therapy. By minimizing the changes in the speech sounds, synthesis artifacts can be minimized too.

To stay close to clinical practice, we asked speech therapists to evaluate the voices on (impressionistic) intelligibility. These therapists already judge the patients speech to apply therapy. So it is natural to let them judge the synthetic manipulations in the same way. Their responses to manipulated stimuli were compared to control stimuli.

Regularizing the source intensity without changing the pitch did improve the perceived intelligibility noticeably. Re-pairing the factor that is normally considered most deleterious, a-periodic or no pitch, did not improve perceived speech quality in this experiment. Replacing the original neo-glottal (LPC) source pitch-synchronously with a completely synthetic neo-ity in this experiment. Replacing the original a-periodic or no pitch, did not improve perceived speech quality noticeably. Replacing the original a-periodic or no pitch, did not improve perceived speech quality noticeably.

The conclusions of this study are therefore, that it is indeed possible to manipulate individual aspects of pathological speech to improve speech quality. It might thus be possible to precede some therapies with synthesizing speech that reflects some aspects of the expected therapy outcome. Based on the projected improvements of the synthetic speech, it might be better possible to balance the relative improvement of the proposed treatment with the efforts needed.

5. Acknowledgements

The authors would like to thank Petra Jongmans, Anne Kornman, Merel Latenstein, Vivienne van de Kamp, Lisette van de Molen, and Maya van Rossum for participating in the experiment and Louis Pols for helpful suggestions and comments. This research was made possible by an unrestricted research grant of Atos Medical, Sweden.

6. References