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: Tracheoesophageal 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 alaryngial 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 denovo 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 connected to a filter that mathematically behaves like a collections of hard tubes. This is not a perfect description of the acoustics of the oral cavity and how it couples with the subglottal air space, e.g., [6, 7, 8]. As a result, there are several practical problems with using the LPC inverse filter signal as a substitute for the voicing source. Even for voiced speech, the LPC analysis cannot determine the resonances of the oral cavity perfectly. These errors will affect the calculated LPC-source signal. Any such problems will be ignored in this study.

The quality of LPC analysis-synthesis speech is not on par 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-emphaisized 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 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 [14, 15]. Before use, all speech recordings are downsampled to 11kHz and overall intensities are normalized. In the current study, the voicing of the speech is labeled by hand by one of us (vS) based on what would have been appropriate for the intended phonemes. For the purpose of synthesizing stimuli in the reported experiment, it was decided to treat the closure of voiced plosives, i.e., the /d/, as unvoiced.

Four classes of synthetic stimuli are used (*with problem area*):

- *AS*: baseline analysis-synthesis with no alterations (controls)
- *EI*: impose a regular voicing amplitude (*amplitude stability*)
- *EP*: impose a regular pitch (*pitch stability*)
- *NS*: fully synthetic source periods (*source spectrum*)

The *AS* stimuli are used as controls. To improve prosody, stress and pitch accents were marked for stimulus types *EI* and *EP*. The accent pattern was chosen to be close to that used in the majority of the readings. Stress markers were placed on the first voiced intervals (syllables) of the words *ook* (also), *weer* (weather), *deze* (this), and *meegewerkt* (contributed), i.e., *OOK het WEER heeft aan DEze tocht MEEgewerkt*.

2.2.1. LPC analysis and (re-)synthesis (AS)

All speech is analyzed with an autocorrelation LPC with 10 poles, equivalent to 5 formants, a window of 25 ms, a step size of 5 ms, and pre-emphasis from 50Hz. Praat scripts used can be found in the supporting files.

An inverse filtering of the downsampled speech recording with the LPC (filter) parameters is performed to obtain the LPC-error signal. At this point the LPC-error signal is de-emphasized (50Hz) and integrated (summed) to obtain a signal that resembles a sound source flow signal. DC and low frequency noise (1/F noise) is removed with a high pass filter (pass Hann band 40-5500Hz). Both the LPC-error signal and the LPC-source can be used as the departure point for manipulating the source signal.

For synthesis, the integrated LPC-source signal is first low-pass filtered (pass Hann band 4-4000Hz) to remove quantization errors, differentiated (sample[i+1] - sample[i]) and normalized, i.e., intensity is set to 70dB, before it is filtered with the LPC-filter. The intensity of the resulting sound signal is set to the required level. Low intensity (silent) and unvoiced intervals are copied from the original recordings and not synthesized.

The original wave form is differentiated and then integrated twice to create a Praat point-process which indicates the position of the source pitch pulses (filtering as above). This procedure is a simplified version of that developed by [16]. Each rising zero crossing can be used to indicate the start of a pitch period. These pulses are then used instead of the normal pitch period marks of Praat.

2.2.2. Stimuli with regularized voicing amplitude (EI)

Voicing amplitude is regularized in two steps. First, the LPC-source is multiplied with the inverse of its Intensity contour (40Hz minimum pitch). Then a triangular raise in the intensity of 3 dB on the 4 stressed syllables is superimposed on an overall decline in intensity from 70-68 dB.

2.2.3. Stimuli with regularized pitch periods (EP)

A regular pitch period is imposed on the re-synthesized speech with a PSOLA technique (Praat Manipulation object). The pitch pulses required by PSOLA are obtained as described above. A pitch contour is defined with a simple declination from 120 to 100 Hz for every speaker. 2 semitone triangular accent peaks are super-imposed on the first voiced intervals of the four accented words.

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 markers is 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

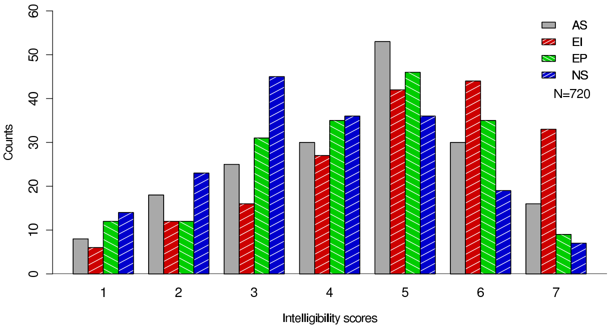

Figure 1: Distribution of intelligibility scores for the baseline analysis-synthesis (*AS*) and manipulated (*EI*, *EP*, *NS*) stimuli. $p < 0.001$ Friedman rank sum test for each stimulus type. N: total number of responses

and one is a phonetician who has collaborated with them. One outside speech therapist participated. Five subjects had experience with speakers used in this experiment. None was aware of the aims of the experiment or the manipulations performed.

The stimuli were presented in an on-line experiment. Subjects had to read an introduction and fill in some personal information. Then the experiment started with 8 practice items. These were the four stimulus manipulations of two versions of the standard sentence. One version was an extra recording of one of the TE speakers. This TE speech version had a very low intelligibility. The other version was spoken by one of the authors (vS) and was an example of intelligible, non-TE, speech. These items were presented in a fixed order of alternating TE speech and non-TE speech stimuli.

The practice items were followed by a pause screen and then the 120 (4x30) stimuli in a pseudo random order, different for each subject. The experiment was self paced and performed at home or at work using headphones. Subjects were not rewarded for their participation.

The task of the subjects was to judge the intelligibility of the (known) utterance on a 7 point scale. As all the subjects were experienced speech therapists or phoneticians, this task was feasible. Subjects heard the stimulus automatically after the web page loaded. After that, they could listen to the stimulus as often as they wanted by clicking a button on the page. Subjects could proceed to the next stimulus after having selected one of the response categories (1-7) by clicking on the “next” button. Attempts to change earlier responses were prevented. Subjects 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*; ν=29, $\chi^2 > 99$, 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 renderings 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 H_0, 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]. It

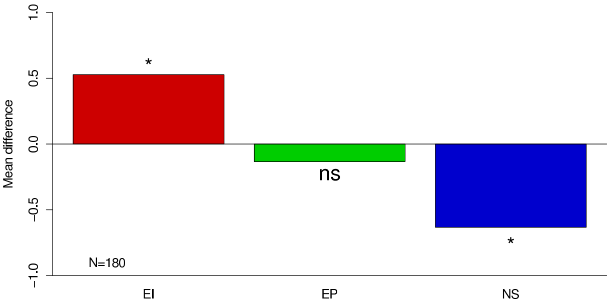

Figure 2: Mean difference in perceived intelligibility scores with respect to plain *AS* baseline stimuli (figure 1). Higher is better quality. *EI*: Equalized Intensity, *EP*: Equalized Pitch, *NS*.: New Source signal. *: $p < 0.001$, ns: $p \geq 0.001$. N: number of responses per stimulus type.

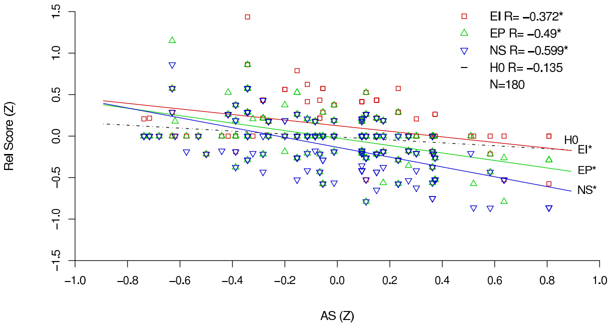

Figure 3: Correlation between difference scores (figure 2) and control *AS* scores (figure 1). Correlations are calculated on the standardized judgements (Z-scores). $H0$: Monte-Carlo simulation of uncorrelated responses. See figure 2 and text.

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. Repairing 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 *glottal* source resulted in the lowest quality. Audible artifacts are very likely with this method and might have affected the results.

It is not yet clear how a change in the voice source intensity contour improves perceived intelligibility. It might be related to the known effects of a hypertonic esophagus. Patients whose esophagus have a high muscle tension have difficulty in producing and controlling sound. The procedure to equalize the source intensity might simulate a loss of hypertonicity or a better sound production in general. If confirmed, such a finding might be relevant for therapeutic interventions.

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

[1] F. J. M. Hilgers, A. H. Ackerstaff, C. J. Van As, A. J. M. Balm, M. W. M. Van den Brekel, and I. B. Tan, “Development and clinical assessment of a heat and moisture exchanger with a multi-magnet automatic tracheostoma valve (provox freehands hme) for vocal and pulmonary rehabilitation after total laryngectomy.” *Acta Oto-laryngologica*, vol. 123, no. 1, pp. 91–99, 2003.

[2] P. Jongmans, “The intelligibility of tracheoesophageal speech: An analytic and rehabilitation study,” Ph.D. dissertation, University of Amsterdam, 2008.

[3] C. J. van As, “Tracheoesophageal speech. a multidimensional assessment of voice quality,” Ph.D. dissertation, University of Amsterdam, Sept. 2001.

[4] T. Haderlein, E. Nöth, H. Toy, A. Batliner, M. Schuster, U. Eysholdt, J. Hornegger, and F. Rosanowski, “Automatic evaluation of prosodic features of tracheoesophageal substitute voice,” *European Archives of Oto-Rhino-Laryngology*, vol. 264, no. 11, pp. 1315–1321, 2007.

[5] M. Moerman, J. Martens, M. Van der Borgt, M. Peleman, M. Gillis, and P. Dejonckere, “Perceptual evaluation of substitution voices: development and evaluation of the (i)infvo rating scale,” *European Archives of Oto-Rhino-Laryngology*, vol. 263, no. 2, pp. 183–187, 2 2006.

[6] K. Matsui, N. Hara, N. Kobayashi, and H. Hirose, “Enhancement of esophageal speech using formant synthesis,” *Acoustical Science and Technology*, vol. 23, no. 2, pp. 69–76, 2002.

[7] S. Fraj, F. Grenez, and J. Schoentgen, “Evaluation of a synthesizer of disordered voices,” in *Proceedings 3rd Advanced Voice Function Assessment International Workshop*, May 2009, pp. 69–72.

[8] ——, “Synthetic hoarse voices: a perceptual evaluation,” in *Proceedings Int. Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications - MAVEBA 2009*, Dec 2009, pp. 95–98.

[9] J. Hanquinet, F. Grenez, and J. Schoentgen, “Synthesis of disordered speech,” in *Proceedings Ninth European Conference on Speech Communication and Technology*, 2005, pp. 1077–1080.

[10] Y. Qi, “Replacing tracheoesophageal voicing sources using LPC synthesis,” *J. Acoust. Soc. of Am.*, vol. 88, pp. 1228–1235, 1990.

[11] P. Jongmans, F. J. M. Hilgers, L. C. W. Pols, and C. J. van As-Brooks, “The intelligibility of tracheoesophageal speech, with an emphasis on the voiced-voiceless distinction.” *Logoped Phoniatr Vocol*, vol. 31, no. 4, pp. 172–181, 2006.

[12] P. Jongmans, T. G. Wempe, H. van Tinteren, F. J. M. Hilgers, L. C. W. Pols, and C. J. van As-Brooks, “Acoustic analysis of the voiced-voiceless distinction in dutch tracheoesophageal speech.” *J Speech Lang Hear Res*, vol. 53, no. 2, pp. 284–297, 2010.

[13] G. Fant, *Acoustic theory of speech production.* Mouton The Hague, 1960.

[14] D. Weenink, “The Klattgrid speech synthesizer,” in *Proceedings of Interspeech 2009*, Sept. 2009, pp. 2059–2062.

[15] P. Boersma and D. Weenink, “Praat: doing phonetics by computer,” Computer program: http://www.Praat.org/, 2009.

[16] B. Yegnanarayana and K. Murty, “Event-based instantaneous fundamental frequency estimation from speech signals,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 614–624, 2009.

[17] R Core Team, “The R project for statistical computing,” Computer program: http://www.r-project.org/, 1998–2010.