USING VOICE CONVERSION AS A PARADIGM FOR ANALYZING BREATHY SINGING VOICES

Karl I. Nordstrom[1]*, Glen A. Rutledge*[2]*, Peter F. Driessen*[1]

[1]University of Victoria
Department of Electrical and Computer Engineering
P.O. Box 3055 STN CSC,Victoria, BC, V8W 3P6, Canada
www.ece.uvic.ca/˜knordstr/, www.ece.uvic.ca/˜peter/
knordstr@ece.uvic.ca, peter@ece.uvic.ca

[2]TC Helicon
Research Department
6710 Bertram Place,Victoria, BC, V8M 1Z6, Canada
http://www.tc-helicon.com/
GRutledge@tc-helicon.com

ABSTRACT

It is generally thought that breathiness can be added to voices by modifying the glottal source within a source-filter model. However, this does not work well when the original voice is very different from the desired breathy voice. In this experiment, a voice conversion algorithm is used to investigate the relationship between the glottal source and the vocal tract filter. The LPC residual from one voice is fed into the LPC filter of another voice. According to a source-filter theory of the voice, the synthesized voice should take on the glottal quality of the LPC source. This hypothesis is evaluated through a perceptual test with a linguistics expert. The results suggest that the vocal tract does have an influence on the perception of breathy voices. Given the narrow nature of this experiment, further testing is recommended to verify these results.

1. INTRODUCTION

Breathiness is typically added to LPC models of the voice by modifying or replacing the LPC residual[1][2]. These techniques focus on modeling the glottal wave and the aspiration noise of the voice. However, these techniques are not very successful when attempting to transform voices that are significantly different from the desired breathy voice. This suggests that the vocal tract may be involved in the perception of breathy voices. A couple of other streams of research support this idea. Recent research in linguistics shows that the lower vocal tract can be influential in the production of different voice qualities[3][4]. Also, others have shown that these kinds of small changes can significantly modify the vocal tract filter[5][6]. In addition, there are situations when acoustic resonances within the vocal tract can influence the glottal source[7]. This influence could also affect the accuracy of the model.

This paper proposes that voice conversion techniques [8][9] can be used to understand particular components of the voice quality without having to model all of the components in detail. This technique can theoretically evaluate the separation of the source and the filter. The point of this evaluation is to determine whether the breathy effect is confined to the LPC residual or whether some components of perceived breathiness are found within the vocal tract.

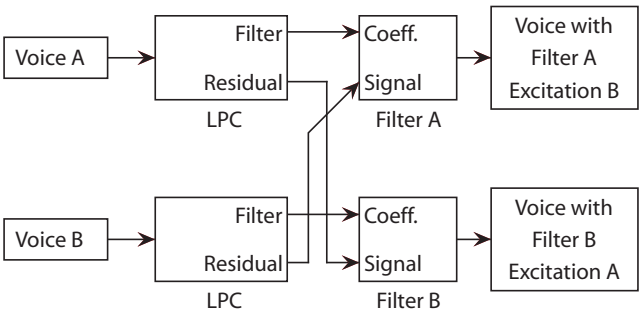

Fig. 1. LPC voice conversion concept.

The basic concept of the voice conversion algorithm is in Figure 1. A breathy and a non-breathy voice sing the same phrase at the same time. The LPC filter is computed for each of these voices in Figure 2. The voices are inverse filtered to extract the residual. The LPC residual from the breathy voice is then fed through the LPC filter from the non-breathy voice. Likewise, the LPC residual from the non-breathy voice is filtered by the LPC filter from breathy voice. Ideally, the result should be that the synthesized voice takes on the glottal characteristics of the LPC residual. The voice that was originally non-breathy should become breathy when it is given a breathy excitation. Likewise, the voice that was originally breathy should become non-breathy when it is given a non-breathy excitation.

2. LPC MODELING

The voice samples used in this experiment were collected from a variety of different sources. Some of them were available from

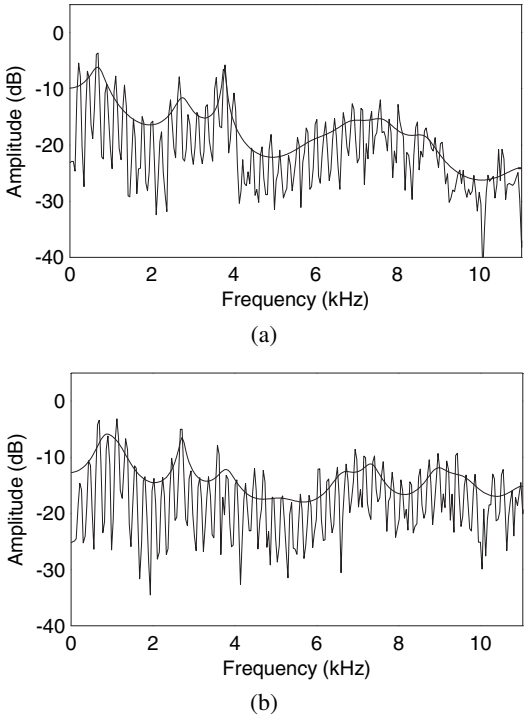

Fig. 2. LPC plot of (a) breathy voice and (b) non-breathy voice. The signal has already been de-tilted.

previous experiments[10] while others were newly recorded. The most ideal samples were those recorded by one person singing or speaking the same vowel with a breathy and non-breathy voice. The voices were recorded at a sample rate of 22050 Hz, which was chosen as a compromise between having high enough frequencies to capture the breathy quality and a low enough sample rate for LPC to work properly.

For these tests, autocorrelation LPC was used. Ideally, covariance LPC provides a more accurate model of the vocal tract if it is calculated while the glottis is closed. However, the glottis does not entirely close for breathy voices and glottal closure is difficult to estimate without additional data.

The LPC algorithm was chosen to have an order of 20 because it corresponds to a typical vocal tract length of 15 cm long[11]. The voice signal is de-tilted with a high-pass filter. This de-tilting compensates for the effect of lip radiation[12]. It also makes the signal more spectrally flat, making it easier for LPC to fit the signal. Theoretically, the LPC residual corresponds to the volume-velocity wave of the glottis if the LPC filter corresponds closely to the vocal tract and the signal is de-tilted before being processed.

The resulting residual is not perfectly flat as seen in Figure 3. Some of the resonances are very peaky and it would help to use bandwidth expansion to make them flatter[13][14]. At the same time, there are some gaps in the spectrum that are not modeled well. LPC does not model zeros well because LPC is an all-pole model. Using bandwidth expansion would make LPC fit the zeros even worse. There is a trade-off here.

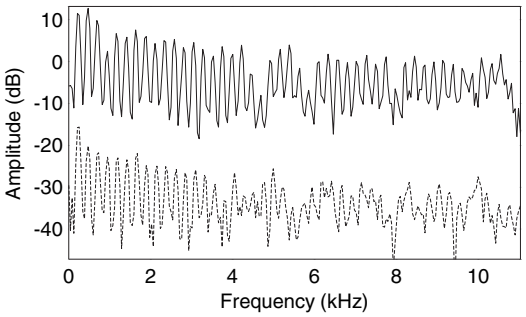

Fig. 3. LPC Residual. The top plot is of a non-breathy voice. The bottom plot is of a breathy voice. An arbitrary gain has been applied for the purpose of the plot. Tilt is included in the residual.

3. PERCEPTUAL TESTING

The results of the experiment were evaluated with the help of a linguistics expert. A preliminary test showed that it was difficult to achieve clear ratings with isolated samples. Therefore, the test was designed to measure the relative difference between a benchmark sample and the other samples. This approach has been used previously with good results for evaluating breathy voices[2]. For each set of four samples, one of the original samples was chosen as a benchmark by which the other corresponding samples were evaluated. The evaluator was not told how each sample was generated or whether the sample was natural or synthesized. The comparison samples were randomized.

The perceptual criteria for this test was drawn from other studies for evaluating breathy voices[1][2]. The parameters from these tests were breathiness, naturalness, vocal effort and nasality. Some other parameters were also added in an attempt to gain a deeper understanding of the perceived configuration of the voice. The parameters are listed below:

- Breathiness:
  (-5 = much less breathy, 0 = no change, 5 = much more breathy)
- Vocal effort: a strained or tense voice quality also known as hyperfunction[1]:
  (-5 = much less vocal effort, 0 = no change, 5 = much more vocal effort)
- Nasality:
  (-5 = much less nasal, 0 = no change, 5 = much more nasal)
- Constriction above the glottis:
  (-5 = much less constriction, 0 = no change, 5 = much more constriction)
- Velarized:
  (-5 = much more velarized, 0 = no change, 5 = much less velarized)
- Creakiness:
  (-5 = much less creaky, 0 = no change, 5 = much more creaky)

Naturalness was evaluated separately, without a benchmark, to get a sense of whether the synthesized samples are close to the original samples in quality. Naturalness was defined as human sounding.

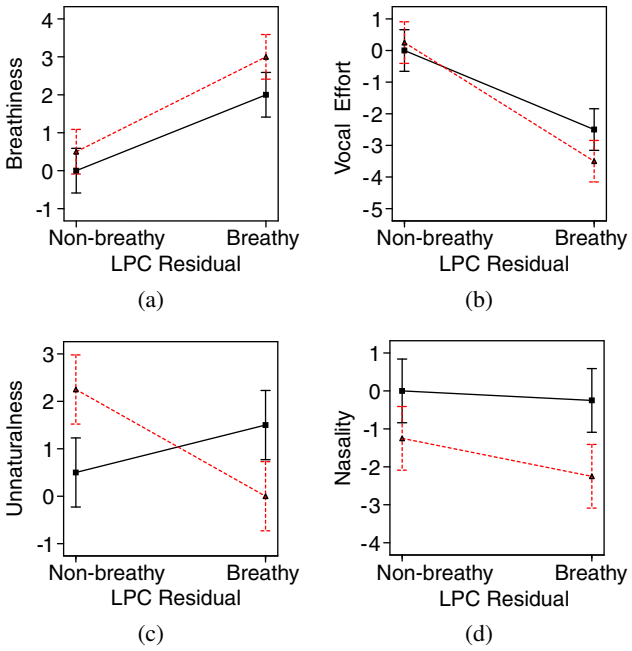

Fig. 4. Interaction plots for (a) perceived breathiness, (b) perceived vocal effort, (c) perceived unnaturalness, and (d) perceived nasality. The dotted and solid lines represent data for the breathy LPC filter and non-breathy LPC filter respectively. The 95% confidence intervals are also plotted. The non-breathy LPC filter with non-breathy LPC residual is at (0,0) because the non-breathy voice was used as a reference sample (except for the rating of unnaturalness).

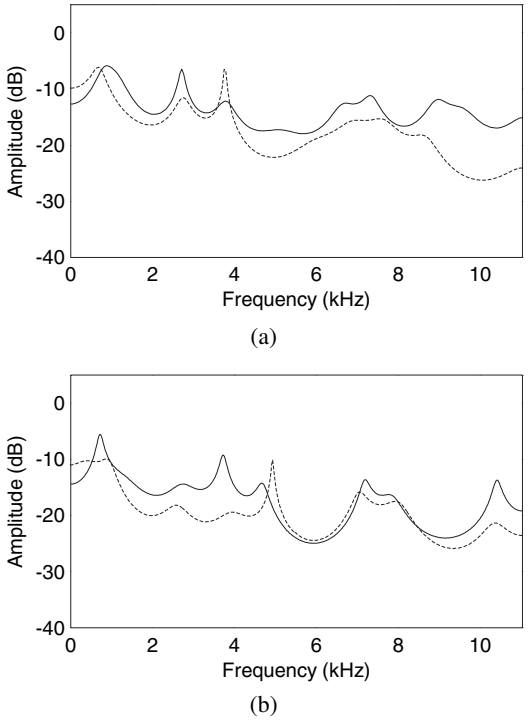

Fig. 5. Vocal tract filters for (a) high pitched male singing /ah/, and (b) low pitched male singing /ah/. Solid line is non-breathy. Dotted line is breathy.

The evaluation was carried out by Dr. John Esling, a professor in linguistics at the University of Victoria. Esling's research investigates different sound production mechanisms within the voice[3][4]. As such, he has a detailed physiological understanding of the voice mechanism and an experienced ear for detecting different voice qualities. The use of an expert listener reduces the risk inherent in the small sample size. However, the test should be repeated with a larger sample size to achieve results that are more broadly applicable.

4. RESULTS

Factorial analysis[15] was carried out on the test results as shown in Figure 4. Constriction and velarization were not statistically significant. The most significant responses were for breathiness, vocal effort, unnaturalness, and creakiness. Creakiness and vocal effort were highly correlated but vocal effort had a larger range. Nasality was rated differently for different vocal tracts and did not change greatly with the excitation Figure 4(d).

The interaction plot for naturalness is found in Figure 4(c). The most obvious thing to observe from this plot is that the original samples sound more natural than the samples with swapped excitations. This is to be expected. However, it also raises the question of whether any unnatural sounds may have been a distraction in the evaluation.

The interaction plot for breathiness in Figure 4(a) shows that there is a large increase in perceived breathiness when the LPC residual from a breathy voice is fed through the LPC filter for a non-breathy voice. We also see that the newly synthesized voice does not achieve the same level of breathiness as the original breathy voice.

A similar phenomenon is seen in the interaction plot for vocal effort but in reverse in Figure 4(b). Vocal effort is negatively correlated with breathiness. When a breathy LPC residual is fed into a non-breathy LPC filter, the perceived vocal effort goes down. Again, the vocal effort does not go all the way to the level of the original breathy voice.

The breathy LPC residual achieves most of the transformation but the transformation is not complete. The LPC filter must account for some of the perceived breathy effect. When we look at the LPC filters we see that there are significant differences between breathy and non-breathy filters, even when the same voice is singing the same vowel at the same pitch (Figure 5 and 6).

An informal experiment was carried out in which an impulse train was fed through the LPC filters. The impression was that the LPC filters from the breathy voices had low vocal effort. The LPC filters from the non-breathy voices sounded like there was high vocal effort. Unfortunately, there was not time for a more controlled listening test.

Some artifacts were present in some of the synthesized data and this may have affected perceptions of breathiness. The voice that was rated the most unnatural was a non-breathy LPC residual fed into a breathy vocal tract. It sounded like a sine wave overlaid on the voice at approximately 500Hz. The frequency of this artifact was confirmed by removing it with a narrow band filter. The artifact was generated by a large resonance in the breathy LPC fil-

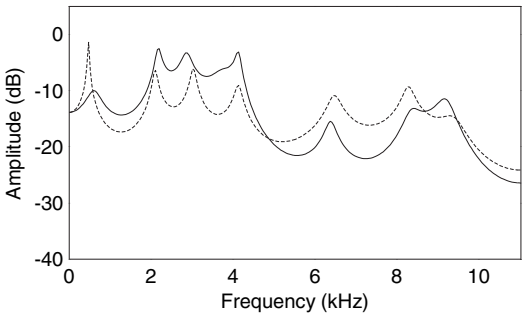

Fig. 6. Vocal tract filters for female singing /ay/. Solid line is non-breathy. Dotted line is breathy.

ter as seen in Figure 6. Bandwidth expansion would help to remove this problem.

5. CONCLUSION

An attempt was made to convert a non-breathy voice into a breathy voice. The LPC filter from a non-breathy voice was excited by the LPC residual from a breathy voice. The resulting voice quality was not as breathy as the original breathy voice. This indicates that the perception of breathiness involves more than the LPC residual. This phenomenon was analyzed with a factorial analysis experiment and the result was found to be consistent. A breathy LPC residual is not capable of fully transforming a non-breathy voice to a breathy voice. In addition, the perception of vocal effort was found to be inversely correlated with breathiness. There was one linguistics expert evaluating the data. The experiment should be repeated with more evaluators to gain greater confidence in the results.

The filters representing breathy and non-breathy vocal tracts were examined and found to be significantly different. An impulse train was used to excite the breathy and non-breathy LPC filters. The result was that the non-breathy LPC filters created the perception of more vocal effort than the breathy LPC filters. This was an informal experiment and the results should be verified in a more controlled way.

Artifacts were present in some of the synthesized voices. This was partially due to peaky resonances in the LPC filters. For clearer results, these artifacts should be removed before repeating the test. Bandwidth expansion is one technique that may be helpful in this regard.

The above algorithm is useful for examining the perceptual influence of different source-filter models. The source-filter models can be investigated without having to explicitly model the glottal pulses and aspiration noise. The greatest opportunity with this technique is to better understand how the vocal tract filter may affect the perception of different voice qualities. In this way we can improve the modeling of breathy voices.

6. REFERENCES

[1] Donald G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, November 1991.

[2] Dennis H. Klatt and Laura C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, February 1990.

[3] John H. Esling, "The laryngeal sphincter as an articulator: How register and phonation interact with vowel quality and tone," in *Western Conference on Linguistics*. November 2002, UBC.

[4] John H. Esling and Jimmy G. Harris, "Expanded taxonomy of states of the glottis," *15th International Congress of Phonetic Sciences*, vol. 1, pp. 1049–1052, 2003.

[5] Ingo R. Titze and Brad H. Story, "Acoustic interactions of the voice source with the lower vocal tract," *Journal of the Acoustical Society of America*, vol. 101, no. 4, pp. 2234–2243, April 1997.

[6] Brad H. Story, Ingo R. Titze, and Eric A. Hoffman, "The relationship of vocal tract shape to three voice qualities," *Journal of the Acoustical Society of America*, vol. 109, no. 4, pp. 1651–1667, April 2001.

[7] Donald G. Childers and Cun-Fan Wong, "Measuring and modeling vocal source-tract interaction," *IEEE transactions on biomedical engineering*, vol. 41, no. 7, pp. 663–671, July 1994.

[8] Donald G. Childers, Ke Wu, D. M. Hicks, and B. Yegnanarayana, "Voice conversion," *Speech Communication*, vol. 8, pp. 147–158, 1989.

[9] D. G. Childers, B. Yegnanarayana, and Ke Wu, "Voice conversion: Factors responsible for quality," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 10, pp. 748–751, April 1985.

[10] Laura Anne Bateman, "Soprano, style and voice quality: Acoustic and laryngographic correlates," M.S. thesis, University of Victoria, 2004.

[11] John D. Markel and Augustine H. Gray, *Linear Prediction of Speech*, Springer-Verlag Berlin Heidelberg, New York, 1976.

[12] Johan Sundberg, *The Science of the Singing Voice*, Northern Illinois University Press, Dekalb, Illinois, 1987.

[13] Peter Kabal, "Ill-conditioning and bandwidth expansion in linear prediction of speech," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. I–824 – I–827, April 2003.

[14] Yoh'ichi Tohkura, Fumitada Itakura, and Shin'ichiro Hashimoto, "Spectral smoothing technique in PARCOR speech analysis-synthesis," *IEEE Transactions on Acoustics, Speech, and Signal Processing (ASSP)*, vol. 26, no. 6, pp. 587–596, December 1978.

[15] R. L. Mason, F. G. Gunst, and J. L. Hess, *Statistical Design and Analysis of Experiments with Applications to Engineering and Science*, John Wiley and Sons, New York, 1989.