Real Time Voice Alteration Based on Linear Prediction

Ping-Fai Yang and Yannis Stylianou

AT&T Laboratories – Research
180 Park Avenue, Florham Park, NJ 07932, USA
[yang, styliano]@research.att.com

ABSTRACT

In this paper we present the application of a set of voice alteration algorithms based on Linear Prediction (LP). We present a study of some potential application areas of voice alteration technology and argue that near real time performance is a critical requirement for many. One benefit of our algorithms is their simplicity and therefore feasibility of implementation in a real time system. To this end, we built an experimental platform on a personal computer. We also present our implementation and user experience from this effort.

1. INTRODUCTION

Voice alteration techniques attempt to transform the speech signal uttered by a given speaker so as to disguise the original voice. It is also possible to modify the original voice to sound like another speaker (the so-called target speaker); this is generally known as voice conversion[1, 2]. In the current study there is no target speaker. However, the output voice can be "designed" to have some specific characteristics.

We see various potential application areas where this technology can be applied, some examples are: by using a voice alteration engine as a speech signal postprocessor to a text-to-speech (TTS) system, voices of tremendous variability can be produced simply and efficiently without the cost incurred for making new recordings. In the entertainment industry, it is often necessary to hire professional voice actors to produce sound tracks for animated characters. A voice alteration system may well replace, or at the very least, enhance the skills involved in sounding like a two foot miniature mouse. User experience in Internet "chat rooms" (places where people can gather and socialize on the Internet using an altered identity called an *avatar* [3, 4]) can be greatly enhanced by a technology for disguising the voice of the speaker. In a similar vein, a lot of newer computer game titles have the ability for interaction with other gamers over the Net. Imagine the ability to chat with someone halfway around the globe while collaborating on fighting alien invaders. The possibilities are just endless.

It is generally assumed that the overall shape of the spectral envelope together with the formant characteristics are the major features controlling speaker identity [5, 6, 7]. While an obvious difference between speakers is the variation in the range of the fundamental frequency (f_0), trying to scale it would give the impression that the same speaker is speaking in a different pitch range. Therefore, it is necessary to modify the spectral information as well: by raising the range of f_0 and shortening the vocal tract results in female or child voice, lowering the range of f_0 while lengthening the vocal tract will give the impression of an adult male voice. It is a well established fact that linear prediction analysis of speech can be used effectively to estimate the system function that models the combined effects of vocal tract response, glottal wave shape, and lips radiation [8]. Modifying this function one can modify the vocal tract response. One major motivation for using linear prediction as the basis of the voice alteration system is its simplicity and therefore ease of implementation as near real time system. We shall see in Section 2 this is a valid requirement for a lot of potential applications for voice alteration technology.

In this paper we present a real time voice alteration system based on the *Linear Prediction* model. Two spectral representations were used: the line spectrum frequencies (LSF) and the pseudo log area ratio (PLAR). The algorithm operates as follows: An LP analysis was performed to produce the aforementioned spectral representation. Then, this set of parameters is modified in order to produce a different linear prediction filter. The new filter is checked for stability and is then excited by the residual signal to retrieve the modified speech signal. As a final step, the energy of the output speech is adjusted to have the same energy as the input signal. The resulting output speech is of high quality while avoiding the problems of spectral discontinuities.

The first part of the paper is devoted to a description of the different applications where the proposed voice alteration system can be used and to the interface of the system. Then, the signal processing techniques used for the current voice alteration system are presented. Next, issues on the real-time implementation of the system are discussed. Conclusions and future work are presented in the last section.

2. APPLICATIONS AND INTERFACE DESIGN ISSUES

While we see various potential applications of voice alteration technologies, we tend to believe that design issues are dictated very much by each particular environment. Let us discuss some examples:

Online Gaming The typical user is not a very sophisticated computer user. Take for example a young child playing on a game console machine, while chatting with her co-players via an Internet connection. The constraint here is to be able to render the altered voice very close to real-time, otherwise the delay will become noticeable to the listeners. Another problem is controlling the parameters in the voice alteration algorithm: think about the knowledge gap one has to bridge in order to explain the underlying functionalities of the parameters that can be tuned in a voice alteration system. One feasible solution is to have several predetermined sets of parameter values that the user can pick from.

Voice Acting The real-time constraint is not as crucial when a person is recording the voice for an animated feature. However, the sound recording engineer will want to exert rather flexible control over the voice output, such as the pitch, duration, and volume level. The challenge here is to come up with an algorithm that is very high quality, yet easy enough to be understood by a technical person who is not necessarily an expert in signal processing.

Assisting hearing impaired persons People with a hearing loss in the high frequency sounds can benefit from a device that can change appropriately the spectral envelope of the speech signal. The constraint on the voice alteration algorithm is of course (near) real-time performance, together with high efficiency to fit into a tiny processor that could possibly fit into a hearing aid.

Text-to-Speech A voice alteration engine can be fitted as a signal postprocessor to a text-to-speech system for providing great variability in the synthesized voice without incurring the cost of making new recordings. Near real-time performance is also critical to maintain good interactivity with the overall system.

Thus we see a range of different forces driving the usability (and therefore acceptance) in these and many other applications of voice alteration. One algorithm, therefore, may not fit all. However, we tend to see a common thread in all these examples: near real-time performance. This led us to examine algorithms that have reasonable quality and can also be implemented efficiently on current hardware platforms. We believe the linear prediction based algorithm presented in this paper does fall in this category.

We also built an experimental software platform on a personal computer running Microsoft Windows. This not only allows us to find out some of the software issues relating to the building of a near real-time application, but also gives a concrete system that our evaluation users can get their hands on. The software was designed such that one can change and experiment with different algorithms during a single execution of the program. While we are still on the topic of interface design issues, we would like to state the observation that in most experimental systems, the researcher/programmer is also a user of the system. Proper design of the program interface will greatly enhance their productivity. Some of our thoughts along this direction will be presented in a later section.

3. SIGNAL PROCESSING ALGORITHMS

Our system is based on LPC modeling of the voice signal. The input speech is split into frames of 30ms duration with an overlap of 15ms. In each frame the LP coefficients are estimated using the Levinson-Durbin algorithm. The residual signal is obtained by filtering the speech samples of the frame with the LP analysis filter. Before extracting the LP coefficients, the signal is pre-emphasized and multiplied with a Hanning window in order to ensure a stable LP analysis filter.

The voice alteration procedure is depicted in Figure 1. Notice that the modification of the excitation signal is shown by a dashed line, as this modification could be done independent of the rest of the other subsystems. For instance, if the above system is used as the back end of a text-to-speech (TTS) synthesis system, the excitation would be modified by the TTS system. Large modifications of the coefficients

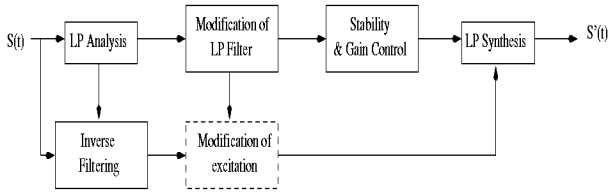

Figure 1: Block diagram of the voice alteration system.

of the LP filter can result in an unstable system. Thus, before driving the system with the excitation signal the filter is checked for stability and its gain is adjusted so that the resulting filtered signal $S'(t)$ matches the energy of the output signal $S(t)$.

An LP analysis filter is given by:

$$A_p(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \quad (1)$$

where p is the order of the LP filter. Representation parameters are then derived from this filter. We tried a number of representation schemes. The first one is the Line Spectral

Frequnecies (LSF), which are the angular positions of the roots of the polynomials [9]:

$$\begin{aligned} P(z) &= A_0 z^p + A_1 z^{(p-1)} + \cdots + A_p \\ Q(z) &= B_0 z^p + B_1 z^{(p-1)} + \cdots + B_p \end{aligned} \quad (2)$$

where

$$\begin{aligned} A_0 &= 1 \\ B_0 &= 1 \\ A_k &= (a_k - a_{p+1-k}) + A_{k-1} \\ B_k &= (a_k + a_{p+1-k}) - B_{k-1} \end{aligned} \quad (3)$$

for $k = 1, \cdots, p$. The reason for selecting LSF to be one of the representational scheme is that these parameters relate closely to formant frequencies and they can be estimated quite reliably. Modifying the roots of the above polynomials, one is essentially changing the formant frequencies of the processed speech signal. Note that the number of control parameters in the case of LSF depends on the order of the LP filter.

The second representation scheme we used was the Pseudo Log Area Ratio (PLAR) parameters[10]. The PLAR parameters are obtained from the LPC reflection coefficients k_i according to the following equation:

$$\begin{aligned} g_0 &= 0 \\ g_i &= g_{i-1} + \log\left\{\frac{1-k_i}{1+k_i}\right\} \end{aligned} \quad (4)$$

with $i = 1, \cdots, p$. Figure. 2 shows the Pseudo Log Area Ratio parameters of a vowel /o/ computed from a segment of a natural speech utterance from a female speaker with g_1 being the front of the vocal tract and g_p being the back of the vocal tract. We can control the size of the vocal tract, first, by changing its length (by shortening or lengthening the front and the back side together or independently) and then computing the new PLAR at the original distances from the front of the vocal tract. Thus, only two control parameters are needed in that case: one modification factor for the front and another for the back of the tract. Compared to the LSF parameters the PLAR are computed faster and their control is easier than for LSFs. Figure. 2 also shows (dashed line) the modified PLAR parameters after a modification of the front by 15% and the back by -15%.

Figure. 3 shows how these modifications of the PLAR parameters result in a modified spectral envelope. Figure. 3 refers to the same vowel as before (/o/) and with the same modification factors for front and back. Comparing the two

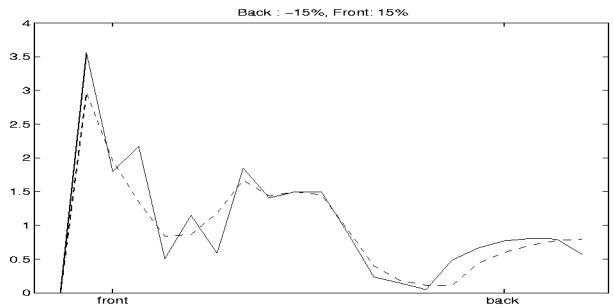

Figure 2: PLAR parameters for a vowel /o/ from a female speaker.

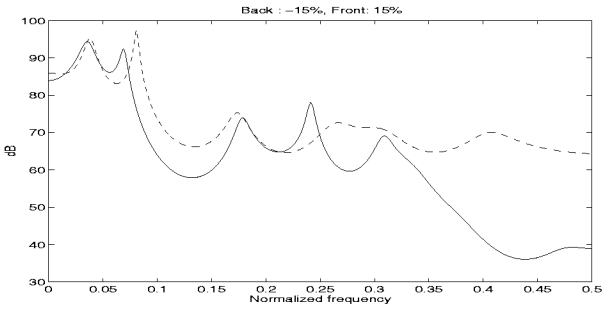

Figure 3: Modifications of the spectral envelope for the vowel /o/. Original (solid line) and modified (dashed line) spectral envelope.

approaches presented here we prefer the second one because of its simplicity in terms of control of vocal tract (only two parameters have to be modified) and of computation load (it's by far easier to compute PLAR than LSF). Both methods provide modified speech of good quality without problems of spectral discontinuities. However, sometimes a slight tonality has been perceived, mostly when we tried to modify unvoiced sounds or voiced fricatives of adult male speakers. This tonality has been removed when a voiced/unvoiced (V/UN) decision was incorporated into the system. This additional feature did not increase the complexity of the system as a simple criterion (based on the first reflection coefficient) has been used. On the other hand, an accurate V/UN decision was not critical for our task. A further improvement has been achieved when the ranges of possible modifications were limited for some frequency bands (mostly between $3000Hz$ and $4000Hz$). This was easier to implement within the LSF approach than within the PLAR.

4. ISSUES ON THE REAL-TIME IMPLEMENTATION OF THE SYSTEM

In order to experiment with the performance of our proposed algorithms, as well as to understand issues in building near real-time programs, we decided to implement an experimental platform on a readily available hardware platform.

The hardware of choice was a sound card equipped PC running Microsoft Windows. Audio input and output data are accessed using the WAV API (Application Programming Interface), though a lower delay API called DirectX became available later. Our experimental platform design includes a single program that controls the user interface, which allows the user to change the voice alternation after startup. Later on we plan to add the capability of incorporating new algorithms into the system at run time by using Dynamic Load Libraries (DLL). The current user program allows for op-

eration in either a live feedback mode or a record-and-play mode, We also would like to have the capability to record and play from saved sound files.

To maintain real-time performance, a careful buffering technique was employed to prevent the system audio buffer from emptying. More precisely, we have to keep rings of buffers for use by the audio input and output devices. When the system has filled and returned to us an input buffer, a copy of it had to be made to internal buffers that are used by the processing algorithm. The returned buffer is immediately fed back to the audio system to prevent buffer exhaustion. Even with our current non-multi-threaded implementation, we found the system to be usable on a 150 MHz Pentium machine running Windows 95.

Another important design goal we imposed was the ease of incorporating new algorithms into the test platform. The voice alteration algorithm can be partitioned into different DLLs that can be installed into the system. Therefore, the speech researcher can reuse our experimental platform in other applications. Moreover, the software is designed using the STRATEGY pattern in [11], meaning that the system uses a C++ class hierarchy for representing various different algorithms. The interface to all these different algorithms is accessed through the base class virtual functions, thereby decoupling the user interface having to know beforehand all possible algorithms that it can provide. By adopting this software architecture, we have removed the need for the person implementing new algorithms from having to learn the user interface code.

We have shown this prototype system to different people and have received generally positive feedback on its usability. We indeed have achieved near real-time performance on near obsolete machines (as of mid-1998). It would be very interesting if we incorporate our voice alteration system into other applications such as an online game to explore user reactions to this technology.

5. CONCLUSION

In this paper we presented a real time voice alteration system based on Linear Prediction (LP) using two spectral representations: the line spectrum frequencies and the pseudo log area ratio. A wide range of voices can be produced by the proposed system. It would be interesting in the future to test the voice alteration system with Internet voice chat applications to find out the acceptance and feasibility of this technology.

Our experience, based on experiments with the real time test program, showed that this technique is of reasonable quality without incurring excessive computational load. Many users were pleased with the range of different output voices produced by our test bed. We have also used the presented voice change algorithm as the back end of a text-to-speech synthesis system. In this arena, it is possible to also alter the gender of the speaker if we also change the fundamental frequency of the speaker.

6. REFERENCES

1. Y. Stylianou, O. Cappé, and E. Moulines, "Statistical methods for voice quality transformation," *Proc. EUROSPEECH*, 1995.
2. M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 655–658, 1988.
3. R. Braham and R. Comerford, "Virtual worlds," *IEEE Spectrum*, Mar. 1997.
4. Blaxxun Interactive, "http://www.blacksun.com."
5. H. Kuwabara and Y. Sagisaka, "Acoustic characteristics of speaker individuality: Control and conversion," *Speech Communication*, vol. 16, pp. 165–173, Feb. 1995.
6. H. Hollien, *The Acoustics of Crime - The New Science of Forensic Phonetics*. Plenum Press, 1990.
7. U. G. Goldstein, "Speaker-identifying features based on formant tracks," *J. Acoust. Soc. Am.*, vol. 59, no. 1, pp. 176–182, 1975.
8. L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, New Jersey 07623: Prentice-Hall, 1978.
9. A. M. Kondoz, *Digital Speech*. John Wiley & Sons, 1994.
10. J. Olive and A. L. Buchsbaum, "Changing voice characteristics in text to speech synthesis," *Technical Memorandum, AT&T Bell-Labs*, July 1987.
11. E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design Patterns*. Addison-Wesley, 1995.