RELATIVE ENERGY AND INTELLIGIBILITY OF TRANSIENT SPEECH INFORMATION

Sungyub Yoo[1], J. Robert Boston[1,2], John D. Durrant[2], Kristie Kovacyk[2], Stacey Karn[2], Susan Shaiman[2], Amro El-Jaroudi[1], Ching-Chung Li[1]

Departments of [1]Electrical Engineering and [2]Communication Science and Disorders, University of Pittsburgh, Pittsburgh, PA 15261, USA

ABSTRACT

Consonants are recognized to dominate higher frequencies of the speech spectrum and to carry more information than vowels, but both demonstrate quasi-steady state (QSS) and transient components, such as vowel to consonant transitions. Fixed filters somewhat separate these effects, but probably not optimally, given diverse words, speakers, and situations, and we suggest that important speech information is contained in transient speech components. To study the relative intelligibility of transient vs. steady-state components, we employed an algorithm based on time-frequency analysis to extract QSS energy from the speech signal, leaving a residual signal of predominantly transient components. Psychometric functions were measured for speech recognition of processed and unprocessed monosyllabic words. The transient components were found to account for approximately 2% of the energy of the original speech, yet were nearly equally intelligible. As hypothesized, the QSS components contained much greater energy while providing significantly less intelligibility.

1. INTRODUCTION

Traditional methods of studying the auditory system and speech intelligibility have emphasized frequency-domain techniques. While it is generally recognized that voicing and steady vowel sounds are largely low frequency and that consonants are dominated by higher frequencies, no single cutoff frequency uniquely separates them. Information on transitions between and within vowel sounds is even more difficult to isolate using fixed-frequency filters. Since speech articulators cannot move instantly from one position to another, initial portions of vowels often show brief frequency shifts that differ among different consonant-vowel (CV) combinations [1]. Conventional vowel-consonant classification and concepts of spectral composition de-emphasize this transition information.

Most human sensory systems are sensitive to abrupt changes in stimuli. We suggest that the auditory system shows the same characteristics and that it is particularly sensitive to time-varying frequency edges, which probably reflect the transition components in speech. Hence, although these transitions represent a small proportion of the total speech energy, they may be critical to speech perception.

The purpose of this project is to isolate and characterize transition information in speech, with the goal of using this information to enhance the intelligibility of speech in noise. This paper describes a method using time-varying filters to identify two speech components: a quasi-steady-state (QSS) component representing primarily vowels and hubs of consonants and a transient component primarily representing transitions between and within vowels. The energy and intelligibility of these two components are compared to the energy and intelligibility of the original speech.

Many investigators have addressed the problem of identifying the start and end of phonemes or word segments for automated speech recognition, but only a few studies have focused specifically on transient components in speech. Yegnanarayana *et al*. proposed an iterative algorithm for decomposition of excitation signals into periodic and aperiodic components to improve the performance of formant synthesis [2]. Zhu and Alwan showed that variable frame-rate speech processing can improve the performance of automated recognition of noisy speech [3]. They used constant duration frames but increased the frame rate when speech models showed that the speech was changing rapidly. Yu and Chan characterized transitional behavior by the onset time and growth rate of each low frequency harmonic component of the transient speech segment [4]. Daudet and Torresani described a method to estimate tonal, transient, and stochastic components in audio signals using a modulated cosine transform and a wavelet transform as a step to improve audio coding [5]. Although these researchers investigated the detection of speech transient information, they did not address the relation of the transient information to speech intelligibility.

Our approach used time-varying bandpass filters (TVBF) to remove predominately steady-state energy from the speech signal [6],[7],[8]. The filters were based on an algorithm described by Rao and Kumaresan, who developed a method to represent a speech signal as a product of components [9]. Section 2 of the paper summarizes the filtering method and explains how the center frequency and bandwidth of the bandpass filters were determined. Psychometric methods to evaluate the intelligibility of transient and original speech are also described. Results presented in Section 3 include relative energy and intelligibility measures of the transient components obtained using mono-syllable words. The implications of the findings as an approach to the enhancement of speech intelligibility are discussed in Section 4.

2. METHODS

Digital speech signals were down-sampled from 44100 Hz to 11025 Hz and highpass filtered at 700 Hz. The low frequency part of the spectrum was removed for reasons discussed later in Methods. This region mostly represents voicing and first formant

information, whereas most of the intelligibility-bearing spectrum of vowels and nearly all consonant spectral power fall above approximately 500 Hz [10]. Because the interest in this study was in speech intelligibility and because the maximum word recognition rate (PB_{max}) of highpass filtered (HPF) speech was the same as the original speech (verified by our experimental measurements of intelligibility as summarized in Fig. 3 and Table II)), we used the HPF speech to begin our analysis.

We assume that the HPF speech is a superposition of QSS and transient components, $x_{HPF}(t) = x_{QSS}(t) + x_{tran}(t)$, where $x_{HPF}(t)$, $x_{QSS}(t)$, and $x_{tran}(t)$ are the HPF speech, QSS, and transient components, respectively. The QSS component is the component that the filter algorithm is intended to remove, and we expect it to include most of the energy in vowels and hubs of consonants. The transient component is the signal that remains after $x_{QSS}(t)$ has been removed.

Three time-varying bandpass filters (TVBF) were used to extract QSS energy from the HPF speech. Each TVBF was implemented as an FIR filter of order 150 with center frequency and bandwidth determined from the output of a tracking filter, which included an all-zero filter (AZF) followed by a single-pole dynamic tracking filter (DTF) [9]. Frequency and amplitude of the tracked component (output of the DTF) were estimated using linear prediction in the spectral domain (LPSD) [9]. The center frequency of each DTF tracked one spectral band of the speech signal. The zeros of the corresponding AZF were set to the frequencies being tracked by the other DTFs to minimize the energy at those frequencies appearing at the DTF input. The pole of the DTF was set to the frequency being tracked by that DTF.

The center frequency of the DTF output determined the TVBF center frequency. The bandwidth of the TVBF was calculated from the speech+noise-to-noise ratio (SNNR) using

$$SNNR = \frac{s(t)}{\sqrt{E[n(t)^2]}}$$

$$BW(t) = \begin{cases} 0 & SNNR < \theta \\ B\left(1 - \frac{\theta}{SNNR}\right) & SNNR \geq \theta \end{cases}$$

where n(t) is a reference noise signal recorded from a silent part of speech, s(t) is the speech+noise signal (AM information from LPSD), B is a maximum bandwidth parameter, and θ is a filter activation threshold [11]. If the amplitude of a tracked component is large, the corresponding TVBF has a wide bandwidth, and if the amplitude is small, the filter has a narrow bandwidth. If the SNNR falls below threshold (SNNR $<\theta$), the bandwidth is set to zero making the filter output zero. The bandwidth increases asymptotically from 0 to the maximum value.

In pilot studies with unfiltered speech, the adaptation of the TVBF was found to be dominated by low-frequency energy. With highpass filtering at 700 Hz., the TVBF were more effective in extracting QSS energy from higher frequencies. As stated earlier, the low frequencies have little influence on intelligibility, and their removal did not affect the intelligibility of the speech. The QSS component was obtained as the sum of the outputs of the three TVBF, and the transient component was obtained by subtracting the QSS component from the HPF speech signal.

Each filter is characterized by two parameters: the maximum bandwidth B and the activation threshold θ, which is the speech+noise-to-noise power at which the filter is activated. The maximum bandwidth must be large enough to capture most of the energy in the spectral band being tracked but small enough to be restricted to a single band. The activation threshold is based on the ratio of speech+noise-to-noise power in a spectral band. It must be small enough to assure that the filter is active during a QSS sound and large enough to not be active during speech transitions.

Pilot tests with a preliminary word set were used to determine the maximum bandwidths and bandwidth thresholds of the TVBF that most effectively removed QSS energy from the HPF speech. The bandwidth parameters were systemically varied between 700 to 1100 Hz and the bandwidth threshold between 5 to 18 dB, and intelligibility of the transient component was assessed qualitatively. A bandwidth threshold of 15 dB and maximum bandwidth of 900 Hz. provided the lowest energy in the resulting transient components while maintaining good intelligibility, and those parameters were used for the results presented here.

To evaluate the relative intelligibility of transient and QSS components, psychometric functions to show growth of intelligibility for original, HPF, transient, and QSS components as signal amplitude increased were determined. Three hundred consonant-vowel-consonant (CVC) words from the NU-6 word lists were processed as described above to provide components for each word [12]. Test words were presented in a quiet background to five volunteer subjects with negative otologic histories and hearing sensitivity of 15 dB HL or better by conventional audiometry (250 – 8 kHz). Subjects sat in a sound-attenuated booth, and test words were delivered manually though headphones. Subjects were asked to repeat the words presented, and the number of errors was recorded by skilled examiners under supervision of coauthor JDD, a certified clinical audiologist. For each component, stimuli were presented at five intensity levels from 0% recognition until recognition did not increase or reached 100%.

Recognition results for each subject (five data points) were fit to an error function, using the nonlinear least-squares fit routine 'lsqcurvefit' (MATLAB, The Mathworks, Inc.). The function minimum was set to zero, and estimates of the maximum (PB_{max} or maximum word recognition rate), midpoint (50% recognition), and slope (measured by the standard deviation parameter of the error function) were obtained. The mean squared difference between the fitted function and the original data divided by the total mean square of the data (R^2) was calculated to assess the adequacy of the fit, with $R^2 > 0.8$ being taken to indicate a satisfactory fit.

The parameters obtained for the original, HPF, transient and QSS versions of the words were tested for significant differences across versions. Because the data were skewed, a Friedman test was used as a non-parametric analysis of variance, followed by Wilcoxon paired comparisons for significant Friedman's results. Parameters were averaged across subjects for summary graphs.

3. RESULTS

An example of decomposition of a speech signal is illustrated in Fig. 1 and 2. The monosyllable word "pike", represented phonetically as /paIk/, spoken by a female speaker was decomposed into QSS and transient components as described above. The original speech, HPF speech, QSS, and

transient components are shown in Fig. 1. The energy in the HPF speech is 16% of the energy in the original speech and the energy in the QSS component is 87% of the energy in the HPF speech (14% of the original speech energy). The QSS component is dominated by the vowel /aI/, from approximately 0.07 to 0.17sec. The remaining 13% of the HPF energy is in the transient component (2% of the original speech energy) which includes energy associated with the noise burst accompanying the articulatory release of /p/ from approximately 0.01 to 0.07 sec., and the articulatory release of /k/ at around 0.38 sec.

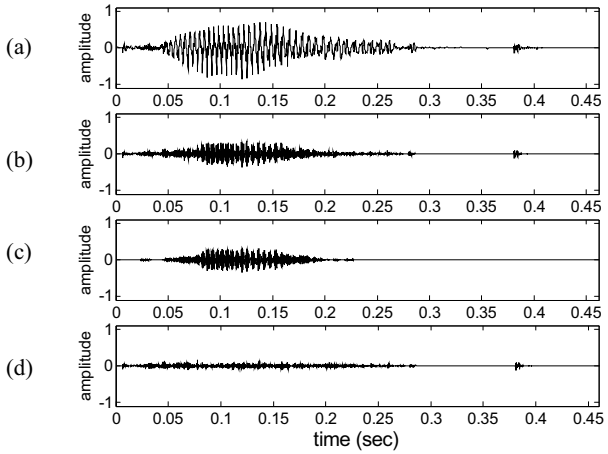

Figure 1. Waveforms of speech signal "pike" : (a) original speech; (b) HPF speech; (c) QSS component; (d) transient component

The sound of the QSS component was very garbled and difficult to identify as the word "pike". On the contrary, the transient component was perceptually similar to the HPF speech, despite having much less energy.

Time-frequency plots of the signal spectra, calculated using a 25 msec. Hamming window, are shown in Fig. 2. The spectrograms are linearly scaled because this scaling shows differences between transient and QSS spectra more clearly than a dB scale, and the original speech spectrum is not shown in this figure because this spectrum is dominated by large low-frequency energy, making it difficult to keep scale levels identical to the other plots. Most of the sustained vowel energy is included in the QSS component, and the transient component primarily includes energy at the beginning and end of the QSS component. In particular, the transient component includes spectral characteristics of both the /p/ and /k/ releases, as well as formant transitions from the /p/ release into the vowel /aI/. The location of the spectral energy in these transients contributes to the perception of place of articulation for both the consonants and the vowel.

When this word was processed with three fixed bandpass filters (center frequencies and bandwidths to best match the center frequencies and bandwidths that were observed in the TVBFs), the sum of the filter outputs (corresponding to the QSS component using the TVBF) contained 95% of the energy in the HPF speech, and it was highly intelligible. The remaining 5% of the signal energy was in the residual component, and it was essentially unintelligible, illustrating that the results obtained depend on the use of time-varying rather than fixed filters.

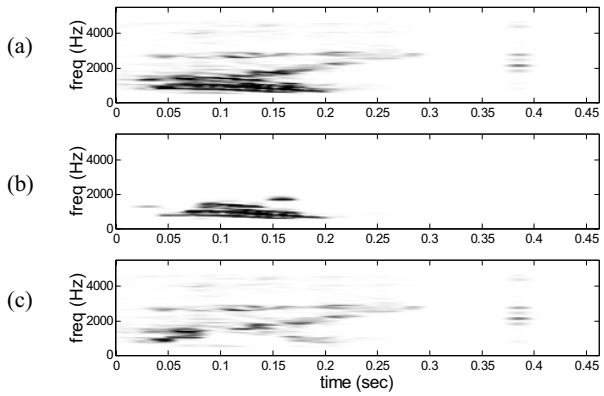

Figure 2. Time-frequency plots of speech components : (a) HPF speech; (b) QSS component; (c) transient component.

These results were typical of all of the words tested. Table I shows the energy in the HPF, transient, and QSS components averaged over the 300 CVC words, as a fraction of the energy in the original speech. The transient components averaged 2% of the original speech energy (18% of the HPF speech energy), and the QSS component averaged 18% of the original speech energy (82% of the HPF speech energy). The QSS component had loudness approximately equal to the HPF speech, but the transient component sounded less loud, as would be expected due to the lower energy.

TABLE I

Mean of energy in the QSS and transient components of mono-syllable words relative to energy in the HPF speech and in the original speech. Standard deviation in parenthesis.

	QSS	Transient
% of HPF speech	82% (6.7)	18% (6.7)
% of original speech	12% (5.5)	2% (0.9)

Each word recognition growth function was based on 50 words, 10 at each of 5 intensities. Each subject was tested with four word versions (original, HPF, QSS, and transient), resulting in each subject listening to 200 words. Word recognition rates were successfully fit to error functions. Of the 20 sets of data (4 different word versions for each of 5 subjects), 18 were fit with $R^2 > 0.9$ and 2 with R^2 between 0.8 and 0.9. The error function parameters for each word version were then averaged across subjects, and a composite growth function for each word version was generated from the averaged parameters. The upper graph in Fig. 3 shows the composite growth functions with unadjusted speech level (the components were tested with relative energies as determined by the algorithm) on the abscissa. HPF was recognized at similar speech levels and had the same PB_{max} as original speech, despite having less energy. The transient component was recognized at higher unadjusted speech levels but had approximately the same PB_{max}. The QSS component was recognized at levels similar to the transient component but had much lower PB_{max}.

The lower graph in Fig. 3 shows growth functions with speech levels adjusted to compensate for the different component energies (that is, 0 dB represents the same energy level in each component.). Original and transient speech were recognized at similar adjusted speech levels (similar absolute energies), while HPF was recognized at lower relative energies.

The QSS speech was recognized at relative speech levels slightly higher than transient speech.

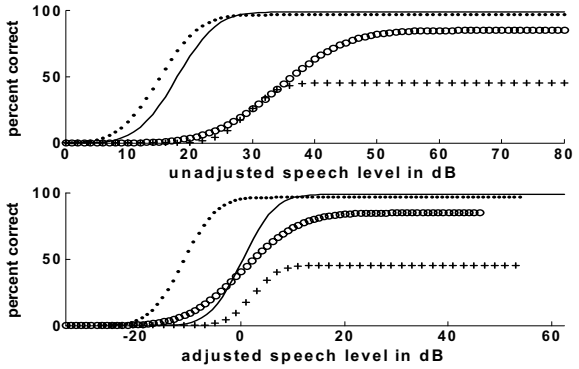

Figure 3 Growth of word recognition based on error function parameters: solid: original speech; dotted: HPF speech; +-+: QSS component; o-o: transient component.

Statistical analyses of the parameters are summarized in Table II. The column 'adj midpt' is the adjusted midpoint values shown in the lower part of Figure 3. The QSS component had a significantly lower PB_{max} than the other components, while the transient component had approximately the same PB_{max} as original and HPF components. The adjusted growth function midpoint of the HPF component was significantly smaller than for original and transient speech, suggesting that this component was detected at lower stimulus levels. The difference in midpoints between QSS and transient speech was not significant, and the slopes of the growth functions showed no significant differences.

TABLE II
Growth function parameters

	Midpoint	adj midpt	PB_{max}	slope
original	17.9 (2.7)	0.3 (2.7)	98.7 (3.0)	7.1 (3.2)
HPF	15.0 (3.8)	-11.2 (3.8)*	96.5 (2.1)	7.2 (2.5)
transient	34.4 (4.6)	0.5 (4.6)	84.9 (14.4)	12.1 (6.3)
QSS	29.2 (11.3)	2.2 (11.3)	45.1 (19.3)*	5.6 (8.5)

* $p < 0.05$ for pair-wise comparisons with other components.

4. DISCUSSION

In order to study the role of transient speech components on speech intelligibility, we implemented a time-varying bandpass filter to extract QSS energy from a speech signal. We refer to the residual signal with low frequency and QSS energy removed as the transient component of speech, and we suggest that it primarily represents transitions between vowels and hubs of consonants. The transient components have approximately 2% of the energy of the original speech but psychometric measures of maximum word intelligibility showed almost equal intelligibility. The QSS components had much greater energy but were significantly less intelligible. We suggest that the QSS component corresponds to speech energy that characterizes sustained vowel sounds and some consonant hubs.

These results suggest that transient components are critical to speech intelligibility, and, if the auditory system is sensitive to transient information, emphasis of the transient components may provide a basis to enhance intelligibility. The transients are expected to be distributed across time and frequency, requiring time-frequency techniques to identify them. The algorithm described here provides one method of extracting predominately transient speech components, and investigations into its utility in enhancing speech intelligibility are currently underway in our laboratory.

5. ACKNOWLEDGEMENTS

The authors would like to express their appreciation to Ken Morton, who prepared the speech material for the psychoacoustic tests. This work was supported by the Office of Naval Research under grant number N000140310277.

6. REFERENCES

[1] A.M. Liberman, P.C. Delattre, *et. al*., "Tempo of frequency change as a cue for distinguishing classes of speech sounds," J. Exp. Psycho., vol. 52, pp. 127-137, 1956.

[2] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components*," IEEE Trans. on Speech and Audio Processing*, vol. 6, pp. 1-11, 1998.

[3] Q. Zhu and A. Alwan*,* "On the use of variable frame rate analysis in speech recognition**,"** *IEEE International Conference on Acoust., Speech, and Signal Processing*, vol. 3, pp. 1783-1786, 2000.

[4] E. Yu and C. Chan, "Phase and transient modeling for harmonic+noise speech coding," *IEEE International Conference on Acoust., Speech, and Signal Processing,* vol. 3, pp. 1467 - 1470, 2000.

[5] L. Daudet and B. Torresani, "Hybrid representations for audiophonic signal encoding," *Signal Processing*, vol. 82, pp. 1595-1617, 2002.

[6] S. Yoo, J. Boston, J. Durrant, A. El-Jaroudi, and C. Li, "Speech decomposition and intelligibility," *Proceedings of the World Congress on Medical Physics and Biomedical Engineering*, August 2003.

[7] J. Boston, S. Yoo, J. Durrant, K. Kovacyk, S. Karn, C. Li, and A. El-Jaroudi, "Relative intelligibility of dynamically extracted transient versus steady-state components of speech," *75th (147th) Meeting of The ASA*, May 2004.

[8] S. Yoo, J. Boston, J. Durrant, K. Kovacyk, S. Karn, S. Shaiman, A. El-Jaroudi, and C. Li, "Relative energy and intelligibility of transient speech components," *EUSIPCO2004*, Sep. 2004.

[9] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Trans. on Speech and Audio Processing*, vol. 8, pp. 240-254, 2000.

[10] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, pp. 1586-1604, Dec. 1979.

[11] M. Li, H. McAllister, N. Black, and T. De Perez, "Perceptual time-frequency subtraction algorithm for noise reduction in hearing aids," *IEEE Trans. on Biomedical Engineering*, vol. 48, pp. 979-988, 2001.

[12] T. Tillman, R. Carhart, "An expanded test for speech discrimination utilizing CNC momosyllabic words," *Northwestern Univ. Auditory Test No 6*, Technical Report, 1966.