RELATIVE ENERGY AND INTELLIGIBILITY OF TRANSIENT SPEECH COMPONENTS

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
Phone: 412 624 3244, fax: 412 624 8003, email:

ABSTRACT

It is generally recognized that consonants are more critical than vowels to speech intelligibility, but we suggest that important information is contained in transient speech components, rather than the quasi-steady-state components of both consonants and vowels. Fixed-frequency filters cannot uniquely separate transients from the more steady-state vowel formants and consonant hubs, even though the former are predominately low frequency and the latter, high frequency. To study the relative speech intelligibility of the transient versus steady-state components, we employed an algorithm based on time-frequency analysis to extract quasi-steady-state 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 quasi-steady-state components contained much greater energy while providing significantly less intelligibility.

1. INTRODUCTION

Most human sensory systems are sensitive to abrupt changes of stimuli. We suggest that the auditory system shows the same characteristics and that it is particularly sensitive to time-varying frequency edges. Most consonants are predominantly brief transients, but some include quasi-steady components, which are also the dominant characteristic of vowels. Since the onset/offset of speech sounds is inherently transient and the consonant-vowel interface is naturally dynamic, the speech signal is replete with transient events. Conventional vowel-consonant classification and concepts of spectral make-up potentially neglect still other transient components. For example, since the articulators cannot move instantly from one position to the other, initial portions of vowel formants show brief frequency shifts that differ among possible CV combinations [1]. Consequently, the transient energy is expressed across the speech frequency range. While the transient components contain a small amount of the total speech energy, compared to quasi-steady state portions of both vowels and consonants, they may be critical to the perception of speech by humans and to machine speech recognition.

Traditional methods of studying the auditory system have emphasized frequency-domain techniques, a perspective that also has dominated concepts of speech intelligibility. 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. Transition information is even more difficult to isolate using fixed-frequency filters as this information is inherently dynamic and can be rather broad band. The purpose of this project was to develop an algorithm to emphasize transient components in speech in order to investigate their role in speech intelligibility.

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 [2]. Their purpose was to improve the performance of formant synthesis. Zhu and Alwan showed that variable frame-rate speech processing can improve the performance of automated recognition of noisy speech [3]. Frame size was constant, but the overlap (rate) was increased 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 speech using a modulated discrete cosine transform and a wavelet transform as a step to improve speech 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 to emphasizing transient information in speech was to use time-varying bandpass filters (TVBF) to remove predominately steady-state energy in the speech signal. 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 [6],[7]. 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 (a surprisingly rare design in the research literature). Results presented in Section 3 include relative energy and intelligibility measures of the transient components (as operationally defined) obtained using mono-syllable words. The implications of the findings are discussed in Section 4, as the results suggest an approach to an efficacious basis for enhancement of speech intelligibility.

2. METHODS

Digital speech signals were down-sampled from 44100 Hz. to 11025 Hz. and then 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 occasionally some first-vowel format information, whereas most of the intelligibility-bearing spectrum of vowels and nearly all consonant spectral power falls above approximately 500 Hz. [8]. Since the interest in this study was in speech intelligibility and the highpass filtered speech was as intelligible as the original speech (verified by our experimental measurements of intelligibility), we used the highpass filtered speech as the reference speech signal.

We assume that the reference speech is a superposition of a quasi-steady-state (QSS) and a transient component, $x(t) = x_{qss}(t) + x_{tran}(t)$, where $x(t)$, $x_{qss}(t)$, and $x_{tran}(t)$ are the reference, quasi-steady-state, 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 $x_{qss}(t)$ have been removed.

Three time-varying bandpass filters (TVBF) were used to extract quasi-steady-state energy from the reference speech. [6],[7] 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). 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 location of the DTF was adjusted by the estimated frequency of the output of that DTF. Then, frequency and amplitude of the tracked component (output of the DTF) were estimated using linear prediction in the spectral domain (LPSD) [6].

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

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

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

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 the maximum bandwidth, and θ is the filter activation threshold [7].

If the amplitude of a tracked component is large, the corresponding DTF has a wide bandwidth, and if the amplitude is small, the filter has a narrow bandwidth. If the SNNR falls in the noise masking region (SNNR $<\theta$), the bandwidth is set to zero to avoid excessive noise energy in the filter output. For larger SNNR, the bandwidth increases asymtotically 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 quasi-steady-state 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 reference 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 reference speech signal.

Each filter is characterized by two parameters: the maximum bandwidth B and the activation threshold θ, which is the speech-power-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 to noise power in a spectral band. It must be small enough to assure that the filter is active during a quasi-steady-state sound and high enough to not be active during speech transitions or noise.

In order to determine the relative intelligibility of the QSS and transient components compared to the original speech, phonetically-balanced CVC words spoken by several male and female speakers (from CDROM CD101R2 from AUDITEC of St. Louis) were used to measure psychometric functions (word recognition vs. stimulus level), providing information on the effective sensation level of the different speech signals tested and the maximum word recognition score (PBmax) for them. Three hundred monosyllable words were divided into 30 sets of 10 words. Each list was processed to generate original, reference (highpass filtered), QSS, and transient component versions.

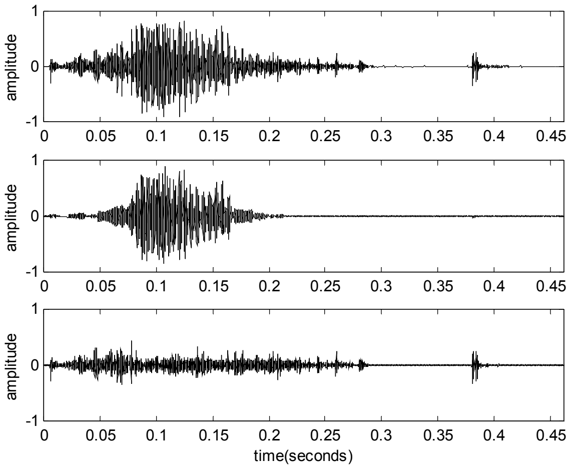

Figure 1. Waveforms of speech signal "pike" spoken by a female speaker: (a) reference speech; (b) QSS component; (c) transient component.

Different sets of words were presented at each of several speech intensities to five young, normal hearing subjects (a robust number by psychoacoustic research standards for such labor-intensive studies). The number of recognition errors was scored by skilled examiners under supervision of coauthor JDD, a certified clinical audiologist. Different lists were used to obtain growth functions for original speech, reference speech, QSS, and transient components to minimize learning or anticipation effects. The test spanned intensity levels from hearing threshold to the maximum recognition level. A maximum recognition score of less than 100% was recorded only if a higher speech amplitude was tested and yielded a recognition score equal to or less than the highest score at a lower level.

The Friedman test was used to perform a non-parametric analysis of variance on the maximum recognition scores for the different components, and post-hoc Wilcoxon's signed-rank tests were used to identify which pairs of scores were different. Results reported here were obtained with words presented in quiet, but essentially the same results were obtained for words presented with various types and intensities of background noise.

3. RESULTS

Our goal was to remove as much energy from the highpass-filtered signal as possible with minimal effect on intelligibility. Pilot tests with the preliminary word set were used to determine the maximum bandwidths and bandwidth thresholds of the TVBF that most effectively removed signal energy from the reference 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 with good intelligibility, and those parameters were used for the results presented here.

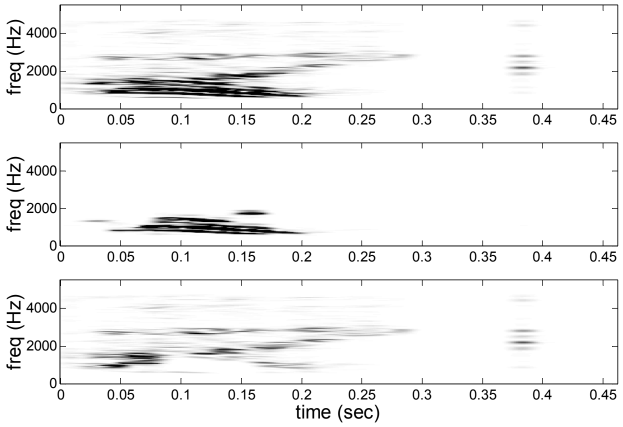

Figure 2. Time-frequency plots of speech components in Fig. 1: (a) reference speech; (b) QSS component; (c) transient component.

An example of decomposition of a speech signal is illustrated in Fig. 1 and 2. A monosyllable word ("pike", represented phonetically as /paIk/) spoken by a female speaker was decomposed into QSS and transient components as described above. The reference, QSS, and transient components are shown in Fig. 1. The energy in the QSS component is 87% of the energy in the reference speech. The QSS component is dominated by the vowel /aI/ in "pike", from approximately 0.05 to 0.17 sec. The remaining 13% of the energy is in the transient component, which includes energy associated with the noise burst accompanying the articulatory release of /p/ from approximately 0.01 to 0.05 sec., and the articulatory release of /k/ at around 0.38 sec.

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

To help visualize the effects of the TVBF, time-frequency plots of the signal spectra were calculated using a 55 msec. Hamming window. Figure 2 demonstrates that most of the sustained vowel energy is included in the QSS component and that the transient component primarily includes energy at the beginning of the dominant components. 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 bandwidth filters (700, 700, and 700 Hz. bandwidths, respectively, to be similar to the 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 highpass filtered 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.

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

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

These results were typical of all of the words tested. Relative energy in the QSS and transient components for the words used to measure the recognition growth functions are summarized in Table I. As expected, most of the speech energy was in the low frequency range that was removed by highpass filtering, and most of the remaining energy was in the QSS component. The energy in the reference speech ranged from 5% to 30% of the energy in the original speech, with an average of 14%. The energy in the transient component ranged from 6% to 43% of the reference speech, with a mean of 18%. The QSS component had loudness approximately equal to the reference speech, but the transient component sounded less loud, as would be expected due to its lower energy.

The maximum recognition scores of words in quiet for the original and reference speech were 100 for all subjects tested. For the transient component, the average score was 92, and for the QSS component, it was 54. The Friedman test was significant (p = 0.014), and the Wilcoxon signed-ranks test showed that only pairs involving QSS were significantly different (p = 0.042).

4. DISCUSSION

In order to study the role of transient speech components on speech intelligibility, we implemented a time-varing bandpass filter to extract quasi-steady-state energy from a speech signal. We refer to the residual signal with low frequency and quasi-steady-state energy removed as the transient component of speech, and we suggest that it includes transitions between vowel formants 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. This intelligibility includes the ability to identify the speaker as well as to distinguish the word being spoken. The QSS components had much greater energy but were significantly less intelligible. They appear to correspond to speech energy that characterizes sustained vowel sounds and some consonant hubs.

These results suggest that transient components are critical to speech intelligibility, and emphasis of the transient components may provide a basis to enhance intelligibility, especially in noisy conditions. 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, L.J. Gerstman, and F.S. Cooper, "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] R. Kumaresan and A. Rao, "Model based approach to envelope and positive instantaneous frequency estimation of signal with speech applications," *J. Acoustic Society of America*, vol. 105, pp. 1912-1924, March 1999.

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

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