R19

A Pitch Synchronous Framework for Processing the Temporal Structure of Speech

Eric Tarr[12], Ashok Krishnamurthy[3], and Susan Nittrouer[1]

[1]Department of Otolaryngology, Ohio State University, [2]Electrical and Computer Engineering, Ohio State University,
[3]Renaissance Computing Institute, University of North Carolina

BACKGROUND

- Cochlear Implants (CI) provide poor spectral resolution which diminishes the perception of frequency-place information.
- Research efforts for CIs have investigated the perception of signal information based on temporal structure.
- Three types of temporal structure have been described: envelope, periodicity, and fine structure. [Rosen, 1992]

Envelope – slow varying modulations related to the syllabic rate.
Periodicity – provides information about the source of excitation.
Fine Structure – informs about the spectrum of a sound, contains the formant pattern, and related to *timbre*.

Purpose : To develop and test a novel approach to manipulating independently the temporal structure of speech.

SPEECH PRODUCTION

Speech can be modeled as a carrier signal and amplitude modulator.

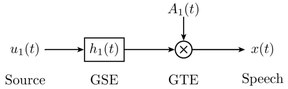
$$x(t) = \mathbb{R}\{A(t)e^{j\phi(t)}\}$$ [Laughlin and Tacer, 1996]

Similarly, speech can be modeled as a source signal and spectral filter.

$$x(t) = h(t) * u(t)$$ [O' Shaughnessey, 1987]

The two models can be combined to represent periodicity, fine structure, and envelope as a source signal, Gross Spectral Envelope (GSE), and Gross Temporal Envelope (GTE).

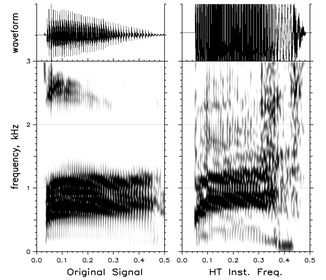
$$x(t) = A_1(t)[h_1(t) * u_1(t)]$$

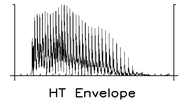


HILBERT TRANSFORM

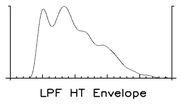
The Hilbert Transform (HT) is one method to represent a carrier signal and a modulating amplitude envelope.

$$\tilde{x}(t) = \mathbb{R}\{a(t)e^{j\phi(t)}\} \quad \phi(t) = \arctan(\frac{\mathcal{H}[x(t)]}{x(t)}) \quad a(t) = |x(t) + j\mathcal{H}[x(t)]|$$



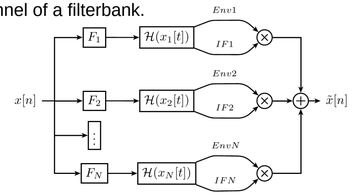


$$\tilde{x}(t) \neq \mathbb{R}\{LPF[a(t)]e^{j\phi(t)}\}$$



AUDITORY CHIMERAS

Smith, Delgutte, and Oxenham [2002] presented Auditory Chimera (AC) processing, a technique to process a signal by finding the Hilbert Transform amplitude envelope and instantaneous frequency signal in each channel of a filterbank.



PITCH SYNCHRONOUS

Pitch synchronous (PS) processing provides an alternative framework for representing the envelope, periodicity, and fine structure of speech.

In voiced portions of speech, each pitch period is detected and processed independently.

GTE, GSE, and f0 are assumed to be constant during each pitch period.

There is one GTE for an utterance and one GSE for each pitch period.

A speech signal can be processed to recover and remove the GTE, GSE, and source signal separately.

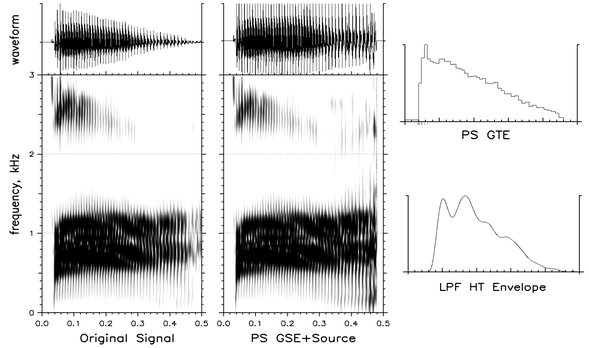
The original signal can be perfectly reconstructed following decomposition.

Additionally, a new speech signal can be reconstructed by replacing one or more of the components.

Pitch Synchronous Temporal Normalization

GTE can be removed from a speech signal by normalizing the amplitude in each pitch period.

GTE can be recovered by measuring the amplitude in each pitch period.



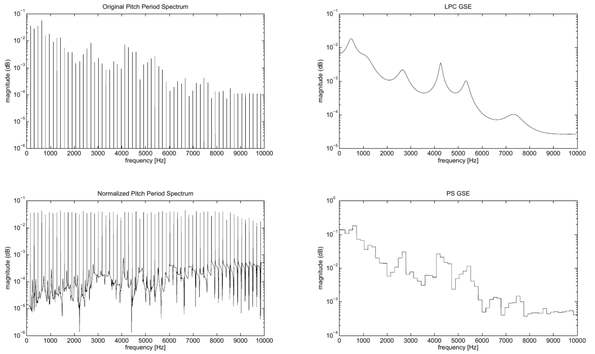
PITCH SYNCHRONOUS

Pitch Synchronous Spectral Normalization

A filterbank with channel cut-off frequencies based on multiples of f0 can be used to process each harmonic in a pitch period separately.

GSE can be removed from a speech signal by normalizing the amplitude of each harmonic in each pitch period.

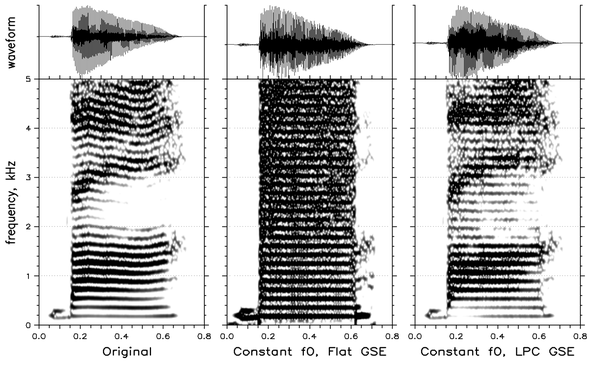
GSE can be recovered by measuring the amplitude of each harmonic within each pitch period.



Pitch Synchronous f0 Normalization

The frequency of f0 can be processed independently by removing the GTE and GSE of the original signal.

Changing the relative length of a pitch period using up-sampling or down-sampling changes f0.



EXPERIMENT

Participants

Six adults listeners with normal hearing (3 male, 3 female).

Stimuli

UCLA CV Database (16 consonants x 3 vowels x 4 speakers).
b-, ch-, d-, f-, g-, j-, k-, m-, n-, p-, s-, sh-, t-, th-, v-, z-
-a, -ee, -oo

Three conditions: Unprocessed, AC processed, PS processed
Interchanging GTE between same syllables

Procedures

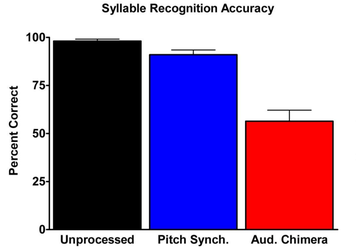
Presentation order of syllable, speaker, and processing condition was randomized for each subject.

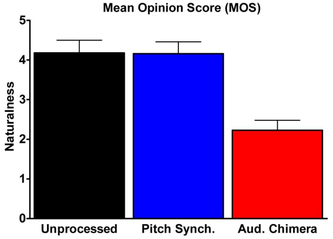
Labeling accuracy and perceived naturalness as a Mean Opinion Score (MOS) on a 5-point Likert scale were recorded for each syllable.

Results

PS processed stimuli were labeled as accurately and rated as sounding as natural as unprocessed stimuli.

AC stimuli were labeled less accurately and rated as sounding less natural.





CONCLUSIONS

PS processing provides an alternative method of representing, recovering, removing, and replacing the temporal structure of speech.

PS processed stimuli can be perceived as sounding natural and can be labeled accurately, an advantage over AC processing.

These signal processing techniques have many applications for perceptual experiments investigating the temporal structure of speech.

ACKNOWLEDGEMENT

Work supported by NIDCD Grant No. DC-000633