Inharmonic Speech: A Tool for the Study of Speech Perception and Separation*

Josh H. McDermott

Center for Neural Science
New York University, USA
jhm@cns.nyu.edu

Daniel P. W. Ellis

Dept. Elec. Eng.
Columbia University, USA
dpwe@ee.columbia.edu

Hideki Kawahara

Faculty of Systems Engineering
Wakayama University, Japan
kawahara@sys.wakayama-u.ac.jp

Abstract

Sounds created by a periodic process have a Fourier representation with *harmonic* structure – i.e., components at multiples of a fundamental frequency. Harmonic frequency relations are a prominent feature of speech and many other natural sounds. Harmonicity is closely related to the perception of pitch and is believed to provide an important acoustic grouping cue underlying sound segregation. Here we introduce a method to manipulate the harmonicity of otherwise natural-sounding speech tokens, providing stimuli with which to study the role of harmonicity in speech perception. Our algorithm utilizes elements of the STRAIGHT framework for speech manipulation and synthesis, in which a recorded speech utterance is decomposed into voiced and unvoiced vocal excitation and vocal tract filtering. Unlike the conventional STRAIGHT method, we model voiced excitation as a combination of time-varying sinusoids. By individually modifying the frequency of each sinusoid, we introduce inharmonic excitation without changing other aspects of the speech signal. The resulting signal remains highly intelligible, and can be used to assess the role of harmonicity in the perception of prosody or in the segregation of speech from mixtures of talkers.

Index Terms: speech synthesis, harmonicity, sound segregation

1. Introduction

Human speech recognition is remarkable for its robustness to background noise. Our ability to recognize speech from mixtures with other sound sources sets humans apart from state-of-the-art speech recognition systems [1], which typically perform well in quiet but are adversely affected by the presence of additional sound sources. The robustness of human recognition to competing sounds reflects our ability to segregate individual sources – to separate the sound energy produced by a target source from that produced by other sources [2].

Human sound segregation relies in part on acoustic grouping cues – sound properties that are characteristic of individual natural sound sources such as speech [3, 4], and that can be used to infer groupings of sound energy from a mixture of sources. Harmonic frequency relations are believed to be among the most powerful of such cues. Harmonicity is the frequency-domain analogue of the periodicity that characterizes many natural sounds, including voiced speech. Periodicity produces frequency components that are multiples of the fundamental frequency (f_0), a relationship known as harmonicity. Frequency components that are harmonically related are generally heard as a single sound with a common pitch, and mistuning a single component of a harmonic series by as little as 1% causes it to be heard as a distinct sound [5]. Moreover, two concurrent tones with different f_0s are typically heard as two distinct sources [6].

Machine systems that attempt to replicate human segregation abilities also make use of harmonicity. Computational auditory scene analysis (CASA) systems typically compute a measure of periodicity and f_0 within local time-frequency cells and then group cells in part based on the consistency of the f_0 estimates. CASA systems in fact rely more strongly on harmonicity than common onset, the other main bottom up grouping cue believed to underlie human segregation [7, 8].

Despite the widespread assumption that harmonicity is critical to sound segregation, its role in the segregation of real-world sounds such as speech remains largely untested. Given the potential importance of spectrotemporal sparsity in the segregation of natural sounds [9, 10], it is conceivable that the most important role of harmonicity could simply be to produce discrete frequency components, the sparsity of which reduces masking and could facilitate common onset and other grouping cues. Moreover, psychoacoustic experiments with artificial stimuli have raised questions about whether harmonicity is in fact critical for segregation. Mistuned frequency components of complex tones can be detected even when the frequencies of all components are increased by a fixed amount, or when the complex is "stretched" such that adjacent com-

*This work was supported in part by the National Science Foundation (NSF) via grant IIS-1117015, by Grants-in-Aid for Scientific Research 22650042 from JSPS, and by the Howard Hughes Medical Institute. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

ponents are no longer separated by a fixed number of Hz [11]. Although such tones are inharmonic (lacking a fundamental frequency common to all the components they contain), component mistuning detection thresholds are comparable to those for harmonic tones. This result suggests that various forms of spectral regularity, rather than harmonicity *per se*, could be most critical to segregation.

The strongest test of harmonicity's importance in segregation would arguably be to compare the segregation of real-world sounds to that of inharmonic equivalents that are matched in other respects. However, speech in particular is nontrivial to manipulate in this manner, as it consists of an interaction of periodic and noise excitation with vocal tract filtering to which humans are exquisitely sensitive. We devised a method to generate inharmonic versions of recorded speech utterances that selectively alters the periodic component of excitation. We used the framework of STRAIGHT, a powerful tool for representing and manipulating speech [12, 13].

2. Methods

Spectral envelopes (used to model vocal tract filtering) were extracted by STRAIGHT from recorded speech and were used to set the amplitudes of constituent time-varying sinusoids (used to model speech excitation). Conventionally, these sinusoidal components would mirror the harmonics of the pitch contour. However, modeling the excitation in this way allows the frequency relations between sinusoids to be manipulated independently of the spectral envelope or the prosodic contour, introducing inharmonicity into an otherwise normal speech signal. This section outlines the original STRAIGHT procedure and its extension to enable inharmonic excitation.

2.1. Original STRAIGHT Framework

Spectral envelope estimation in STRAIGHT consists of a two-stage procedure to eliminate interference from periodic speech excitation [13]. In the first stage, temporal interference is eliminated by averaging power spectra calculated at two time points separated by half a pitch period. In the second stage, spectral interference is eliminated by spectral smoothing using an f_0-adaptive rectangular smoother, followed by post-processing to preserve harmonic component levels based on consistent sampling theory. These frequency domain procedures are implemented with cepstral liftering. More details are provided in [14].

Excitation estimation in STRAIGHT also relies on a temporally stable representation of the power spectrum and combines this with a temporally stable representation of instantaneous frequency [15]. Excitation is represented using a time-varying fundamental frequency $f_0(t)$ (for the voiced, deterministic component of excitation) and time-varying parameters to describe colored noise (for the unvoiced, random component of excitation). The original STRAIGHT framework synthesizes voiced excitation with a sequence of pulses, with each pulse being the minimum phase impulse response of the estimated vocal tract filter at that time point. Fractional pitch control is implemented with a linear phase shifter. The voiced and unvoiced components are combined using a sigmoid function (defined by the boundary frequency between the voiced and unvoiced components, and a transition slope) [16].

2.2. Sinusoidal Modeling of Voiced Excitation

To permit the manipulation of harmonicity, the pulse-based voicing synthesis of the original STRAIGHT procedure was replaced by a sum of multiple sinusoids. Our implementation extends a previous instantiation of sinusoidal excitation modeling in STRAIGHT [17].

Let $A(t, f)$ represent the amplitude at a time-frequency location (t, f) of the spectral envelope estimated using the STRAIGHT procedure. The deterministic (voiced) component $s(t)$ of the sinusoidal synthesis procedure can be defined by the following equation:

$$s(t) = \sum_{n=1}^{N(t)} A(t, f_n(t)) \cos\left(2\pi \int_0^t f_n(\tau) d\tau + \varphi_n\right) \quad (1)$$

where $f_n(t)$ represents the time-varying frequency of the n-th constituent sinusoid and φ_n represents its initial phase (set to zero for the experiments described here). The total number of harmonic components $N(t)$ at time t is adaptively adjusted to keep the highest component frequency below the Nyquist frequency.

Instead of directly implementing Equation 1, as in [17], we approximate it here using a time-varying filter and a fixed frame-rate overlap-add procedure (3 ms frame rate and 50% overlap between adjacent Hanning-windowed frames). A linear phase FIR filter, derived from $A(t, f)$, is applied to each frame using a 1024 sample (64 ms) FFT buffer. This is essentially a cross synthesis VOCODER framework, whose minimal restrictions on the input signal make it straightforward to vary the excitation. To synthesize speech, $s(t)$ is added to the unvoiced speech component estimated as in conventional STRAIGHT.

2.3. Inharmonicity Manipulations

Following manipulations from prior psychoacoustics studies [11, 18], we altered the frequencies of speech harmonics in three ways.

- Shifting: the frequencies of all harmonics were increased by a fixed proportion of the f_0, preserving the regular spacing (in Hz) between components. Hence, the frequency of harmonic n became:

$$f_n(t) = nf_0(t) + af_0(t) \quad (2)$$

- Stretching: the frequency spacing between adjacent components was increased with increasing component number:

$$f_n(t) = nf_0(t) + bn(n-1)f_0(t) \qquad (3)$$

- Jittering: a distinct random offset (uniformly distributed between -30% and +30% of the f_0) was added to the frequency of each component:

$$f_n(t) = nf_0(t) + c_n f_0(t) \qquad (4)$$

For comparison we also synthesized a substitute for whispered speech, in which a low amplitude noise component (26 dB below the level of the voiced component of the regular synthetic speech, an amount that sounded fairly natural to the authors) was added to the usual unvoiced component in lieu of sinusoidally modeled voicing.

3. Results

Figure 1 displays spectrograms of one original speech utterance and the synthetic variants that resulted from our synthesis algorithm. It is visually apparent that the synthetic harmonic rendition is acoustically similar to the original, as intended. The inharmonic variants, in contrast, deviate in their spectral detail. Close inspection reveals that the frequencies of the shifted inharmonic version are translated upwards by a small amount in frequency (such that the component frequencies are no longer integer multiples of the component spacing). The stretched and jittered versions lack the regular spacing found in harmonic spectra, while the simulated whisper lacks discrete frequency components. In all cases, however, the coarse spectro-temporal envelope of the original signal is preserved by virtue of STRAIGHT's speech decomposition, and the unvoiced components in the original speech, which are processed separately, are reconstructed without modification. All synthetic renditions remain highly intelligible, as can be confirmed by listening to the demos available online: `http://labrosa.ee.columbia.edu/projects/inharmonic/`.

Although much of the acoustic structure needed for speech recognition remains intact, the inharmonicity that results from the synthesis is readily audible. Unlike the synthetic renditions that preserve harmonicity, the inharmonic versions do not sound fully natural, perhaps due to weaker fusion of frequency components and/or the absence of a clear pitch during voiced speech segments. To quantify the physical basis of this effect, we used Praat to measure the instantaneous periodicity of each type of synthetic signal for a large number of speech utterances from the TIMIT database. As shown in Figure 2, the periodicity histograms for both the original recordings and their synthetic harmonic counterparts have a steep peak near 1, corresponding to moments of periodic voicing. In contrast, all three types of inharmonic signals lack strong periodicity despite the presence of discrete frequency components. The simulated whisper synthetic speech also lacks periodicity, as expected from noise excitation.

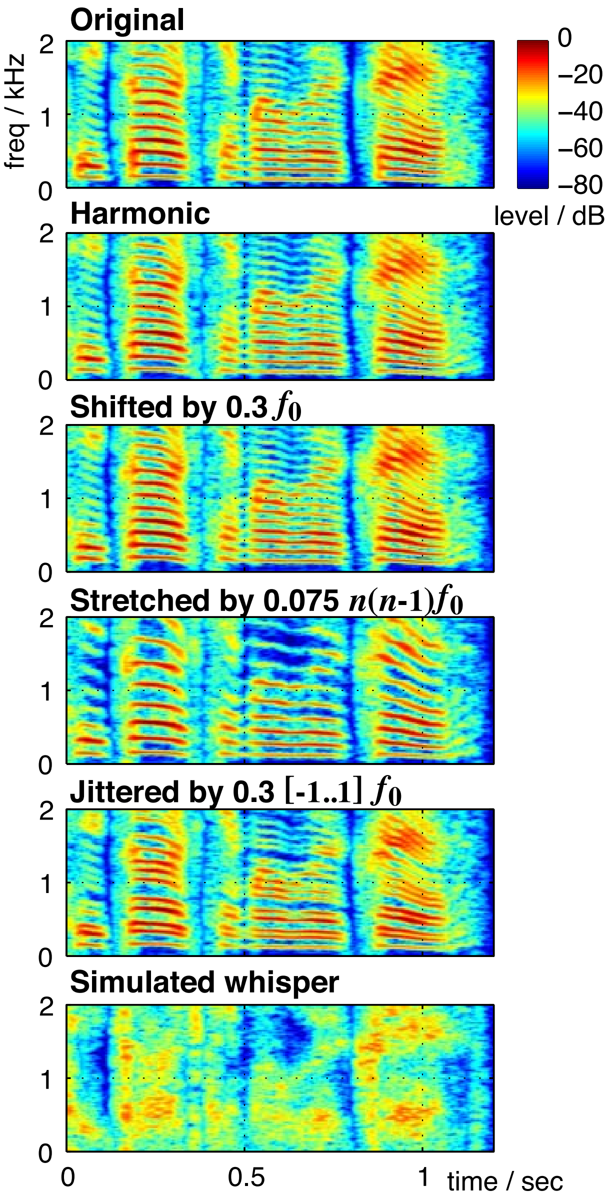

Figure 1: Spectrograms of original and modified tokens of the utterance ”Two cars came over a crest”. The frequency axis extends only to 2 kHz to facilitate inspection of individual frequency components.

Although the individual frequency components of the inharmonic speech utterances trace out the same contour shape as the components of the harmonic speech, the absence of periodicity impairs the extraction of an f_0 contour (in this case by Praat), as shown in Figure 3. The f_0 track for the harmonic synthetic version closely mirrors that of the original (note the overlap between blue

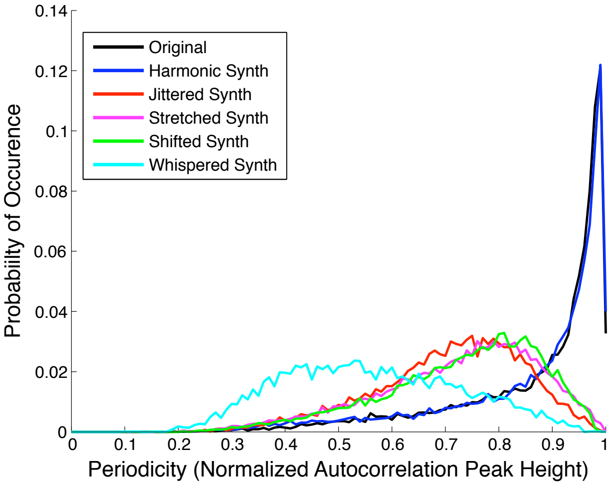

Figure 2: Histograms of instantaneous periodicity for recordings of speech utterances and different synthetic renditions thereof. Data obtained from 76 randomly selected sentences from the TIMIT database.

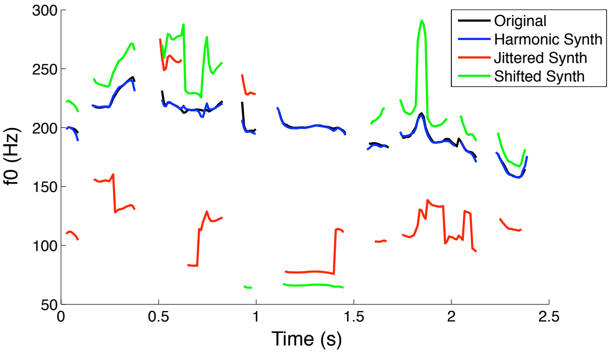

Figure 3: f_0 tracks extracted from an example original speech utterance and its synthetic variants. The stretched inharmonic version is omitted for visual clarity.

and black), but the inharmonic variants do not. Our subjective observations suggest that many aspects of prosody are nonetheless preserved in inharmonic speech, a topic that will be interesting to explore experimentally.

The most exciting application of inharmonic speech stimuli may be to the study of sound segregation. We informally compared the ease of hearing a target speaker mixed with competing talkers, for harmonic, inharmonic, and whispered synthetic speech. Although definitive conclusions will require formal measurements over a large corpus, our subjective impression was that harmonic speech was somewhat easier to perceive in a mixture than was inharmonic speech, with whispered speech noticeably more difficult than inharmonic. In some cases it seemed that harmonic speech derived an advantage from its pitch contour, which helps to sequentially group parts of speech.

4. Conclusions

Inharmonic speech utterances can be synthesized using a modification of the STRAIGHT framework. They are intelligible but lack a clear pitch and sound less fused than veridical harmonic speech. Inharmonic speech signals may be useful for the study of prosody and speech segregation.

5. References

[1] R.P. Lippmann, “Speech recognition by machines and humans,” *Speech Comm.*, vol. 1, no. 22, pp. 1–15, 1997.

[2] A.S. Bregman, *Auditory Scene Analysis*, Bradford Books, MIT Press, 1990.

[3] M. Cooke and D.P.W. Ellis, “The auditory organization of speech and other sources in listeners and computational models,” *Speech Comm.*, vol. 35, no. 3, pp. 141–177, 2001.

[4] J.H. McDermott, “The cocktail party problem,” *Current Biology*, vol. 19, no. 22, pp. R1024–R1027, 2009.

[5] B.C.J. Moore, B.R. Glasberg, and R.W. Peters, “Thresholds for hearing mistuned partials as separate tones in harmonic complexes,” *J. Acoust. Soc. Am.*, vol. 80, pp. 479–483, 1986.

[6] C. Micheyl and A.J. Oxenham, “Pitch, harmonicity, and concurrent sound segregation: Psychoacoustical and neurophysiological findings,” *Hearing Research*, vol. 266, no. 1-2, pp. 36–51, 2010.

[7] G.J. Brown and M. Cooke, “Computational auditory scene analysis,” *Comp. Speech and Lang.*, vol. 8, no. 4, pp. 297–336, 1994.

[8] D.L. Wang and G.J. Brown, “Separation of speech from interfering sounds based on oscillatory correlation,” *IEEE Tr. Neural Networks*, vol. 10, no. 3, pp. 684–697, 1999.

[9] M. Cooke, “A glimpsing model of speech perception in noise,” *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, March 2006.

[10] D. P. W. Ellis, “Model-based scene analysis,” in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. Wang and G. Brown, Eds., chapter 4, pp. 115–146. Wiley/IEEE Press, 2006.

[11] B. Roberts and J.M. Brunstrom, “Perceptual segregation and pitch shifts of mistuned components in harmonic complexes and in regular inharmonic complexes,” *J. Acoust. Soc. Am.*, vol. 104, pp. 2326–2338, 1998.

[12] H. Kawahara, “STRAIGHT, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds,” *Acoustical Sci. and Tech.*, vol. 27, no. 6, pp. 349–353, 2006.

[13] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, “TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation,” in *Proc. IEEE ICASSP*, 2008, pp. 3933–3936.

[14] H. Kawahara and M. Morise, “Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework,” *SADHANA*, vol. 36, no. 5, pp. 713–722, 2011.

[15] H. Kawahara, T. Irino, and M. Morise, “An interference-free representation of instantaneous frequency of periodic signals and its application to F0 extraction,” in *Proc. IEEE ICASSP*, 2011, pp. 5420–5423.

[16] H. Kawahara, M. Morise, T. Takahashi, H. Banno, R. Nisimura, and T. Irino, “Simplification and extension of non-periodic excitation source representations for high-quality speech manipulation systems,” in *Proc. Interspeech2010*, 2010, pp. 38–41.

[17] Hideki Kawahara, Hideki Banno, Toshio Irino, and Parham Zolfaghari, “Algorithm AMALGAM: Morphing waveform based methods, sinuisoidal models and STRAIGHT,” in *Proc. IEEE ICASSP*, 2004, pp. 13–16.

[18] J.H. McDermott, A.J. Lehr, and A.J. Oxenham, “Individual differences reveal the basis of consonance,” *Current Biology*, vol. 20, no. 11, pp. 1035–1041, 2010.