Harmonic Alternatives to Sine-Wave Speech

László Tóth, András Kocsor

Research Group on Artificial Intelligence
Hungarian Academy of Sciences, Szeged, Hungary
{tothl,kocsor}@inf.u-szeged.hu

Abstract

Sine-wave speech (SWS) is a three-tone replica of speech, conventionally created by matching each constituent sinusoid in amplitude and frequency with the corresponding vocal tract resonance (formant). We propose an alternative technique where we take a high-quality multicomponent sinusoidal representation and decimate this model so that there are only three components per frame. In contrast to SWS, the resulting signal contains only components that were present in the original signal. Consequently it preserves the harmonic fine structure of voiced speech. Perceptual studies indicate that this signal is judged more natural and intelligible than SWS. Furthermore, its tonal artifacts can mostly be eliminated by the introduction of only a few additional components, which leads to an intriguing speculation about grouping issues.

1. Introduction

Among the numerous perceptual experiments with spectrally reduced speech, probably the most perplexing is sine-wave speech (SWS) [1]. In SWS experiments a sum of three time-varying sinusoids is generated, each of them mimicking in amplitude and frequency the corresponding speech formants. When asked to listen 'in speech mode', many subjects are able to transcribe SWS surprisingly well. Unprepared listeners, however, report hearing only chirps, whistles or computer bleeps. An obvious explanation is that the sine waves of SWS are not necessarily harmonic and thus do not have a common fundamental, which is, perceptually, a very important characteristic of natural (voiced) speech. In this paper we examine methods for creating stimuli that are similar to SWS in the sense that they also consist of only a couple of sinusoids at a time, but these components preserve harmonicity. Moreover, because these algorithms are based on spectral reduction, the resulting signals consist solely of frequency components that were present in the original signal. We expected these representations to sound more natural than SWS, and we expected increased intelligibility because harmonicity is an important grouping cue as well.

2. Sine-Wave Speech

All the SWS signals used in this study were generated with the Praat software [2] running the SWS script of Chris Darwin [3]. This algorithm estimates the formant frequencies using LPC. Formant amplitudes are then picked from a wideband FFT spectrum. Finally, these frame-by-frame estimates are smoothed to get continuous curves and remove certain types of artifacts. The narrowband spectrum and the SWS replica of a speech excerpt are shown in Fig. 1a (ORIG) and 1b (SWS). Clearly, the main differences between natural and SWS speech are that:

- SWS lacks the fine structure of (voiced) speech, that is, the modulation of the envelope by the pitch, which manifests itself as horizontal lines on the narrowband spectrogram. Moreover, the three sinusoid curves of SWS do not correspond to any of the real frequency components of natural speech and in general are not harmonic, so they do not have a common fundamental.
- The peaks of SWS do not resemble natural speech formants as the latter, in sharp contrast to the former, have a broadband structure.
- In SWS the slow changes characteristic of formants are present in the signal components themselves, while in natural speech formants are present only implicitly as changes in the spectral envelope. Thus while the continuity of the SWS components may be helpful in tracking them, it is also highly unnatural at the same time.

Our goal was to create alternatives to SWS that also consist of only a couple of sinusoids at a given time, but these sinusoids preserve the original harmonic structure of natural (voiced) speech. The first technique we tried reintroduces pitch harmonics into SWS, while the other two methods sought to select proper harmonics from the original signal spectrum itself.

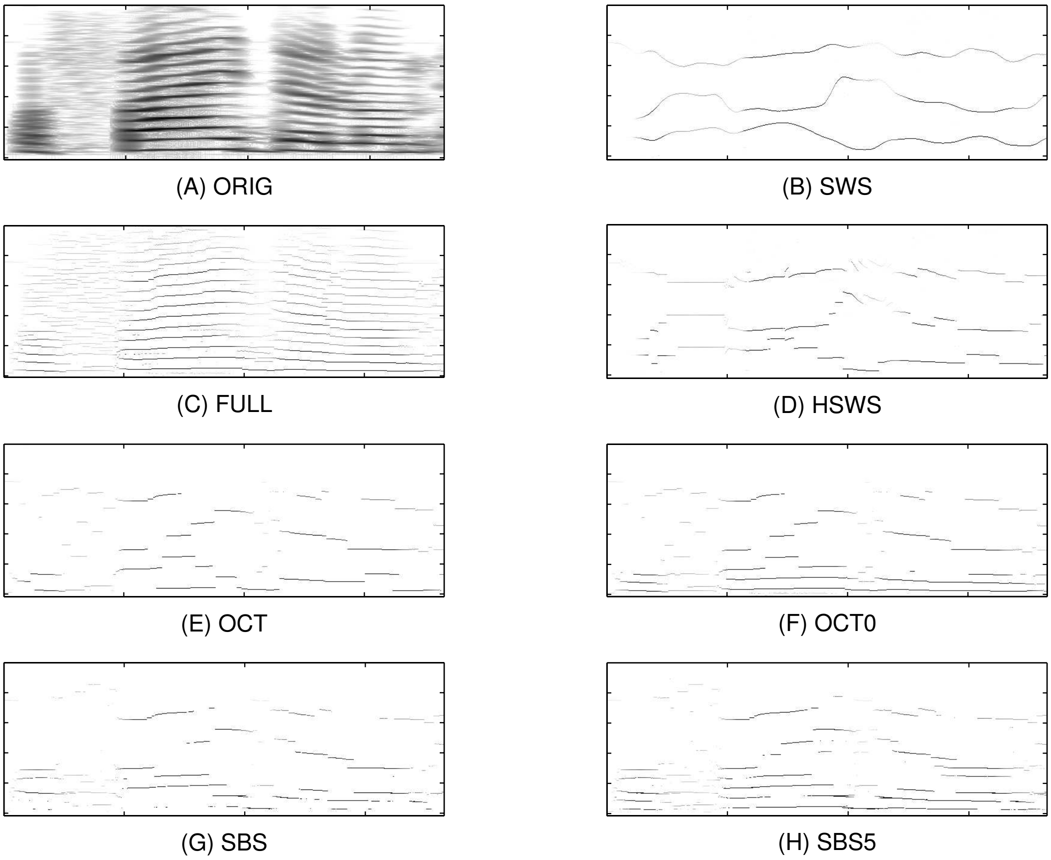
3. SWS with Harmonics Reinserted

Our first idea was to modify the original SWS algorithm so that the sinusoidal frequencies are always rounded to the nearest integer multiple of the pitch. Implementing this, of course, required a pitch estimation of the original signal. It was performed with a conventional autocorrelation-based routine with the pitch of unvoiced frames set to the pitch of the last frame that was judged voiced. Finally, the resulting pitch curve was smoothed with a simple 1-pole filter.

The resulting sinusoidal tracks are shown in Figure 1d (HSWS). Notice that the continuity of the components is no longer preserved, but the spectrum is now composed of short harmonic tracks. However, the spectral envelope of this signal is the same as that for the original SWS.

4. Sinusoidal Models with Decimated Components

Due to its algorithmic constraints SWS is clearly a very weak model of the spectral envelope, so the above 'harmonic insertion' technique will introduce spectral components that were not present in the original signal. An obvious alternative is to start from a model that already represents speech as a combination of time-varying sinusoids. Then experiments with spectral reduction can be performed by decimating the model components.


(A) ORIG

(B) SWS

(C) FULL

(D) HSWS

(E) OCT

(F) OCT0

(G) SBS

(H) SBS5

Figure 1: *Sinusoidal components of the test stimuli. ORIG: narrowband spectrogram of a speech excerpt; SWS: Sine-wave speech; FULL: Sinusoidal representation with all components retained; HSWS: Sine-wave speech with reinserted harmonics; OCT: Reduction based on octave dominance with the lowest filters fused; OCT0: Reduction based on octave dominance with no restriction on the filters; SBS: Reduction using the SBS model with 3 components kept; SBS5: Reduction using the SBS model with 5 components kept.*

4.1. The Sinusoidal Model

The ‘Deterministic Plus Stochastic’ (also known as ‘Harmonic Plus Noise’) Model represents sound signals as a sum of time-varying sinusoids with a stochastic or ‘noise’ component added [4, 5]. The simplest algorithm for the estimation of the model parameters consists of the following steps:

- A wideband spectrogram is calculated. Presuming that the harmonics have been resolved, an initial estimate on the number of sinusoidal components, their amplitude, frequency and phase may be obtained by peak-picking.
- The frame-based estimates are refined by tracking the peak trajectories. This helps ensure the coherence of the component parameters across frames. The result are a set of sinusoid tracks that ‘come alive’, continue for a number of frames, and then die.
- The residue signal is regarded as inharmonic noise and is usually modelled by means of filtered white noise.

In our experiments we simplified two points of the model. Firstly, to preserve the similarity to SWS, we omitted the noise component. We found that this introduced no serious artifacts because the model did very well in describing unvoiced phones with very short inharmonic tracks. Secondly, we ignored the original phases and took care only to keep the component phases coherent across frame boundaries. Like McAulay and Quatieri [5] we found that although ”the resulting speech was perceived as being different in quality from the original speech ... it was very intelligible and free of artifacts”. Thus we ascertained that the artifacts arising in the experiments were due to spectral reduction and not any inadequacy of the model itself (for the sinusoidal spectrum see Fig. 1c (FULL)).

4.2. Spectral Reduction Based on Octave Dominance

Evidently, SWS and the simplified sinusoidal model are very similar. The only difference is that SWS is restricted to three continuous components. With the goal of creating harmonic signals we have already had to abandon continuity, so all that remains for us is to constrain the model so that it only has three active sinusoids at a time. This of course requires proper strategies for selecting the phonetically most important components.

To perform something analogous to formant extraction in SWS, we tried out several metrics in order to select those components that are ‘locally dominant’. Defining ‘locality’ using critical bands would have led to too many components, so we needed wider bands. There is psychoacoustic evidence that speech features are distributed over spectral bands as wide as one octave [6]. In vowel perception, formant integration over 3.5-4 Bark wide bands has been observed [7]. Thus we centered octave-wide rectangular windows on each frequency bin, and a given bin was retained only if it had the highest amplitude in the given window. Because these octave-wide filters resolve the first 1-4 harmonics, the filter bandwidth was restricted to a minimum of 400 Hz. With this modification we found that the algorithm preserves 2-4 components per frame in general, and so is comparable to SWS as regards data reduction. Figures 1e (OCT) and 1f (OCT0) show what remained of the original spectrum using this technique, with and without the 400 Hz minimum bandwidth restriction.

4.3. Spectral Reduction with the In-Synchrony-Bands Spectrum

Although the octave bandwidth of the previous approach was based on psychoacoustic observations, the technique itself is rather ad hoc. We looked for alternatives that are more firmly established in psychoacoustics or neurophysiology - something like a (simple) auditory model. The In-Synchrony-Bands-Spectrum (SBS) model of Ghitza was found to be relevant to our experiments [8]. In this scheme the sound signal is decomposed via an auditory filter bank, and each filter votes on the strongest or 'dominant' component in its output signal. Spectral reduction can then be easily performed by retaining only those components whose 'dominance counter' is above some threshold.

Ghitza himself observed that he could get very good quality speech with only 10 components [8]. The low bit-rate speech coder of Wan et al is built on the same technique and retains only 8 spectral lines per frame [9]. Even with these dramatic reductions both authors reported very good intelligibility along with some tonal artifacts. However, to be comparable with SWS, we had to go one step further and restrict the number of components to just 3. The resulting spectrum can be seen in Fig. 1g (SBS). We found both the spectrum and the synthesized speech very similar to the result of the octave dominance technique. Furthermore, with 5 components we got a result very similar to the octave dominance method with no restriction on the filters (see Fig. 1h (SBS5)).

5. Listening Tests

The main conclusion about sine-wave speech is that speech intelligibility and quality are not necessarily related. With this in mind, we designed separate listening tests to judge the intelligibility and naturalness of our stimuli. In addition, a third experiment was carried out to assess whether the signals were speech-like or not. The test subjects were university students, all unfamiliar with SWS and speech perception experiments in general. The test sentence was chosen quasi-randomly from the first large Hungarian speech corpus [10]. This means that we took the first sentence that was relatively short, of good quality and contained no hesitation or other kind of pronunciation error. From the sentence chosen six stimuli were generated, namely sine-wave speech (SWS), sine-wave speech with reinserted harmonics (HSWS), a spectrally reduced sentence based on octave dominance (OCT0), the same with filter bandwith restricted to at least 400 Hz (OCT), and a spectrally reduced sentence using SBS with 3 (SBS) and 5 (SBS5) components preserved.

In Experiment (A), 8-8 test persons listened to each stimulus. They were told nothing about the stimulus and were asked to identify what they heard. Our aim was to see whether they found the stimulus speech-like or not.

In Experiment (B) subjects were told that they were going to hear a sentence that had undergone some kind of special distortion, and they were asked to transcribe the sentence as precisely as they could. 12-12 subjects listened to each stimulus and they were allowed 4 listenings, with the requirement that they put down a guess after each of them.

Finally, in Experiment (C) subjects had to assess the naturalness of the stimuli. To aid their judgement, we included the original sentence in the test, and subjects were asked to listen to each possible pair of the seven stimuli. They had to assign each pair to a scale with the following categories: 'much less natural', 'considerably less natural', 'somewhat less natural', 'similar', 'somewhat more natural', 'considerably more natural', 'much more natural'. These scores were quantified by the values 1/7, 1/5, 1/3, 1, 3, 5, 7, respectively. The pairwise scores were then averaged to obtain a triangular comparison matrix. It was converted into a full matrix by inserting 1s into the diagonal and filling the lower triangle with the reciprocals of the upper triangle elements (exploiting the antisymmetry of the pairwise relation).

6. Results

Table 1 summarizes how unprepared listeners identified the stimuli. As can be seen, the harmonic stimuli behaved no better than SWS. Only a small fraction of the subjects realized that they had to do with speech, their typical replies being 'reversed speech', or 'fast-forward speech'. The most common reply however was 'science-fiction sounds'. On the other hand, it turned out that the two stimuli with the additional components (OCT0 and SBS5) were always and undoubtedly judged to be speech (most subjects in fact immediately recognized and transcribed the sentence as well).

Table 1: *Number of speech-related guesses (out of 8).*

SWS	HSWS	OCT	SBS	OCT0	SBS5
2	4	3	3	8	8

Table 2 shows the average number of syllables correctly recognized after 1, 2, 3 and 4 listenings, respectively. Apparently, harmonicity did not help improve the intelligibility of SWS. The two spectral reduction techniques did significantly better, but were still far from perfect. The increasing number of hits with more listenings clearly shows that, similar to SWS, they are highly unnatural and require severe adaptation. However, somewhat surprisingly, the additional components were enough to render them perfectly intelligible. Practically no recognition error was made with OCT0 and SBS5.

Table 2: *Average number of recognized syllables (out of 13).*

listenings	SWS	HSWS	OCT	SBS	OCT0	SBS5
1	0.00	0.33	0.00	0.41	12.83	13.00
2	0.33	0.33	2.41	5.00	13.00	13.00
3	1.08	0.76	3.75	6.25	13.00	13.00
4	1.41	1.16	4.50	7.00	13.00	13.00

Lastly, Table 3 shows the average pairwise comparison matrix obtained in the naturalness quality tests. The matrix was evaluated using the Analytic Hierarchy Process (AHP) technique, which is a general tool for multiple criteria decision making problems [11]. The output of the method is a weighted preference list $w_1, ..., w_n$ associated with the alternatives. It is the eigenvector corresponding to the largest eigenvalue of the pair-

Table 3: *The average comparison matrix.*

	ORIG	SWS	HSWS	OCT	SBS	OCT0	SBS5
ORIG	1.00	0.14	0.17	0.18	0.17	0.32	0.32
SWS	6.99	1.00	1.63	4.03	2.60	4.95	5.20
HSWS	5.65	0.61	1.00	1.15	1.80	4.40	5.00
OCT	5.46	0.24	0.86	1.00	1.05	4.00	3.20
SBS	5.65	0.38	0.55	0.95	1.00	3.95	3.60
OCT0	3.12	0.20	0.22	0.25	0.25	1.00	1.00
SBS5	3.12	0.19	0.20	0.31	0.27	1.00	1.00

wise comparison matrix. In our case this largest eigenvalue was $\lambda_{max} = 7.2627$, which indicates that the subjects voted very consistently: the usual consistency index $(\lambda_{max} - n)/(n - 1)$ gives a rather small value, namely 0.0438. The weight vector itself is shown in Fig. 2, normalized such that the original sentence equals 100%.

7. Discussion

When asked about their impressions of SWS, many subjects describe it as if the talker had some terrible laryngeal deformity. That is, they attribute its oddness to the missing fine structure of glottal pulses. Hence it was more than reasonable to expect that adding harmonicity would dramatically increase naturalness. Although it indeed did so to a certain extent, its effect was much smaller than we had expected. Especially disappointing was the result that the harmonic signals still do not have the impression of speech on naive listeners. When looking for an explanation, we quickly realized that all three harmonic signals suffer from the same type of artifact. Namely, subjects reported hearing the speech signal being mixed with trains of short whistles. They judge this less unnatural than SWS only because they consider an additive tonal noise more acceptable than a distortion inherent in the speech production process. These tonal artifacts seem to be responsible for the diminished intelligibility of the stimuli as well.

Where do the artifacts come from? Because we did not insert any extra components in the signal, the only possible source is that certain components separate out from the speech complex and form an independent perceptual stream. Probably the main reason for this is that, to introduce harmonicity, we had to sacrifice the continuity of the components. Even worse, both spectral reduction criteria work on a local, per-frame basis, and ignore the evolution of the component trajectories. Consequently they break the continuity of the tracks and create false onset and offset cues. The rise of the tonal artifacts thus supports the view that harmonicity is a weaker grouping cue than common onset and offset.

The dramatic increase in both quality and intelligibility after the addition of the extra trajectories (see stimuli OCT0 and SBS5) can also be explained by grouping processes, because these trajectories are at very low frequencies and so carry only minimal phonetic information. However they present a strong cue for the detection of F0 continuity, and thus greatly help integrate the sinusoid tracks into one coherent speech stream.

8. Conclusion and Future Work

We found that the sinusoidal plus noise model combined with component decimation offers a viable alternative to creating spectrally reduced speech-like stimuli in a way similar to sine-wave speech. Our experiments showed that these signals are more intelligible and judged more natural than sine-wave speech. On the other hand, these stimuli have their own special type of artifact, which we suppose to be attributable to the fake onset and offset cues introduced via the spectral reduction algorithms. This is strongly supported by the fact that the addition of some extra components - practically the first two harmonics - dramatically reduces these artifacts. In future work we plan to conduct more experiments on trying to understand what the exact conditions are when the coherence of the speech signals breaks up and the tonal artifacts arise. We also plan to modify the spectral reduction algorithm so that it will work with whole trajectories instead of just frame-based peaks.

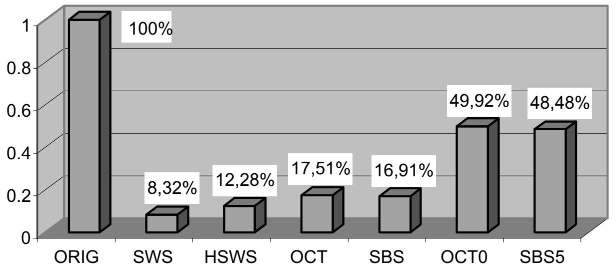


Figure 2: *Naturalness of the stimuli, relative to the original.*

9. References

[1] Remez, E. R., Rubin, P. E., Pisoni, D. B. and Carrell, T. D., "Speech Perception Without Traditional Speech Cues", Science, Vol. 212, 1981, pp. 947-950

[2] Boersma, P. and Weenink, D., "Praat, a System for Doing Phonetics by Computer", <http://www.praat.org>

[3] <http://www.biols.susx.ac.uk/home/Chris_Darwin/>

[4] Serra, X. and Smith, J., "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition", Computer Music J., 14(4), pp. 12-24, 1990.

[5] Quatieri, T. F. and McAulay, R. J., "Audio Signal Processing Based on Sinusodial Analysis/Synthesis", in: Applications of Digital Signal Processing to Audio and Acoustics (ed. Kahrs, M. and Brandenburg, K.), Kluwer, 1998

[6] Allen, J. B., "How Do Humans Process and Recognize Speech?", IEEE Trans. on SAP., Vol. 2., No. 4., Oct. 1994, pp. 567-577

[7] Chistovich, L. A. and Lublinskaja, V. V., "The center of gravity effect in vowel spectra and critical distance between the formants", Hear. Res., 1, pp. 185-195, 1979.

[8] Ghitza, O., "Auditory Nerve Representation Criteria for Speech Analysis/Synthesis", IEEE Trans. on ASSP, Vol. 35, No. 6., June 1987, pp. 736-740

[9] Wan, W., Au, O. C., Keung, C. L. and Yim, C. H., "A Novel Approach of Low Bit-Rate Speech Coding Based On Sinusoidal Representation and Auditory Model", Proc. Eurospeech'99, Budapest, Hungary, 1999, pp. 1555-1558.

[10] Vicsi, K., Tóth, L., Kocsor, A., Gordos, G. and Csirik, J., "MTBA - A Hungarian Telephone Speech Database", Híradástechnika, Vol. LVII., No. 8, 2002. (in Hungarian)

[11] Saaty, T. L., The Analytic Hierarchy Process, McGraw-Hill, New York, 1980.