Enhancement of Reverberant and Noisy Speech by Extending Its Coherence

Scott Wisdom*, Thomas Powers*, Les Atlas*, and James Pitton†*

*Department of Electrical Engineering, University of Washington, Seattle, USA

†Applied Physics Laboratory, University of Washington, Seattle, USA

Main contribution

- We propose a novel speech enhancement algorithm for removing reverberation and noise from recorded speech data.
- Compared to conventional methods, our approach results in:
  - Substantial improvement in PESQ and other objective metrics.
  - Fewer artifacts in informal listening.
- Our method effectively increases the analysis window duration that can be used for voiced speech.
  - We extend the *coherence time*, which is the duration over which an analysis method is coherent with the signal.
  - Conventional methods assume a speech coherence time of 10-30 ms; we extend this time to 128 ms.

System block diagram

Input is $\mathbf{M}$ channels of reverberant, noisy speech:

$\mathbf{y_m[n] = h_m[n] * s[n] + v[n]}$.

Output is estimate of clean speech: $\mathbf{\hat{s}[n]}$

$\mathbf{y_{1:M}[n]}$ → Beamforming: $\mathbf{M=1}$: Do nothing; $\mathbf{M=2}$: Delay-and-sum; $\mathbf{M=8}$: MVDR → STFChT* (vs STFT) → Suppress reverberation and noise using Habets MMSE-LSA [1] enhancement → Inverse STFChT* (or STFT) → $\mathbf{\hat{s}[n]}$

$\mathbf{y_{1:M}[n]}$ → Blindly estimate $\mathbf{T_{60}}$ → Suppress reverberation and noise using Habets MMSE-LSA [1] enhancement

* *"Short-time fan-chirp transform", see panels B and C.*

A. Blindly estimate T_{60}

- Suppress additive noise in each channel using Ephraim and Malah MMSE-LSA [2] and concatenate enhanced channels.
- Use maximum-likelihood blind $\mathbf{T_{60}}$ estimator by Löllmann et al. [3].
- $\mathbf{T_{60}}$ estimation improves with more data (i.e., more channels).

B. Short-time fan-chirp transform (STFChT) [4]

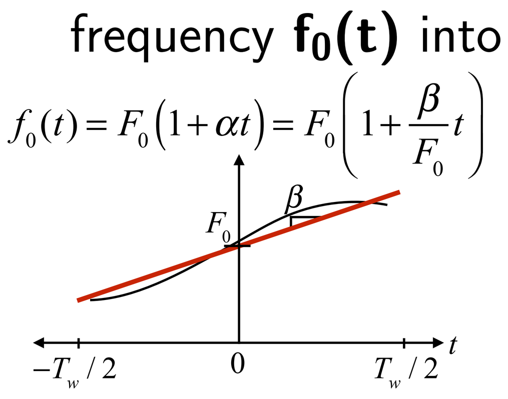
The STFChT is used to implement our method of extending the coherence time of analysis. Definition of STFChT:

$$\mathbf{X_d(f,\hat{\alpha}_d)} = \int_{-T_w/2}^{T_w/2} w(\tau) x_d(\phi_{\hat{\alpha}_d}^{-1}(\tau)) e^{-j2\pi f\tau} d\tau \qquad (1)$$

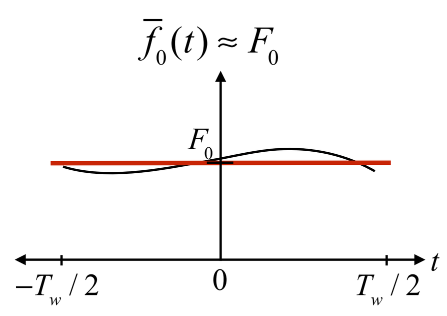
- $\mathbf{w(t)}$ is an analysis window of duration $\mathbf{T_w}$.
- $\mathbf{x_d(t) = x(t - dT_{hop})}$, $\mathbf{t \in [0, T_w]}$, is a short frame of a time-domain signal.
- $\phi_{\hat{\alpha}_d}(\mathbf{t})$ is a linear phase trajectory
- $\phi_{\hat{\alpha}_d}^{-1}(\mathbf{t})$ is a time-warping function.
- $\hat{\alpha}_\mathbf{d}$ is an estimated chirp rate α for the $\mathbf{d}$th frame.

$\mathbf{x_d[n]}$ → Estimate $\hat{\alpha}_d$ → Time-warping using $\phi_{\hat{\alpha}_d}^{-1}(\mathbf{n/f_s})$ → $\mathbf{w[n]}$ → FFT → $\mathbf{X_d(k,\hat{\alpha}_d)}$

- Time-warping converts signals with linearly time-varying fundamental frequency $\mathbf{f_0(t)}$ into signals with approximately constant $\mathbf{f_0(t)}$.



→ Time-warping using $\phi_{\hat{\alpha}_d}^{-1}(\mathbf{n/f_s})$ →

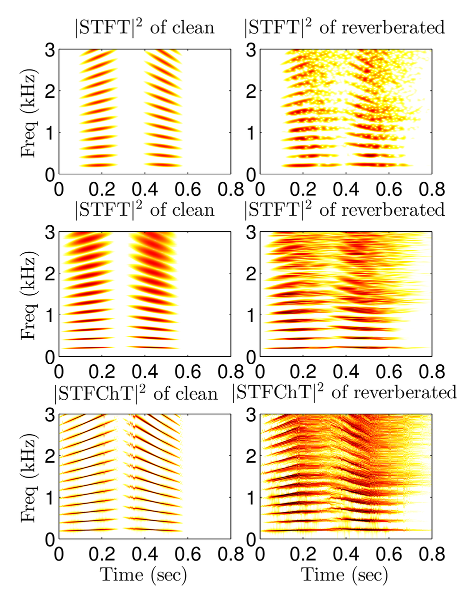


C. Inverse STFChT

- Time-warping implemented as combination of oversampling and interpolation, which achieves almost perfect STFChT reconstruction.

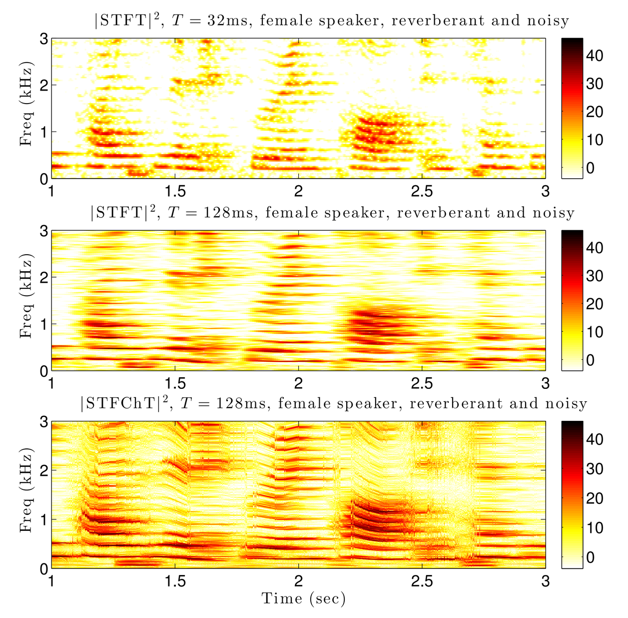
$\mathbf{X_d(k,\hat{\alpha}_d)}$ → IFFT → $\mathbf{w^{-1}[n]}$ → Time-warping using $\phi_{\hat{\alpha}_d}(\mathbf{n/f_s})$ → $\mathbf{\hat{x}_d[n]}$

D. STFT versus STFChT (32 ms and 128 ms window durations)



Synthetic harmonics.

A direct demonstration of the advantage of extending coherence: examples of STFT with 32 ms and 128 ms windows versus STFChT with 128 ms window.

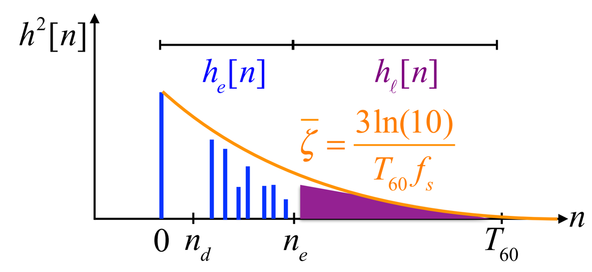


Reverberated speech utterance.

E. Suppress reverberation and noise using Habets MMSE-LSA [1]

- Employs a statistical model of reverberation:

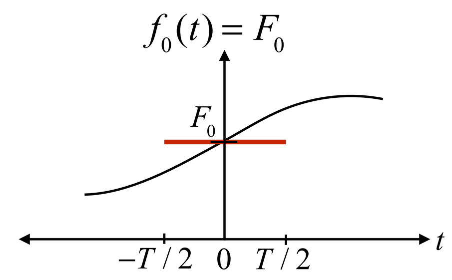
$$\mathbf{E\left[h^2[n]\right]} = \begin{cases} \sigma_d^2 e^{-2\bar{\zeta}n}, & 0 \le n < n_d \\ \sigma_r^2 e^{-2\bar{\zeta}n}, & n \ge n_d \\ 0 & \text{otherwise.} \end{cases}$$

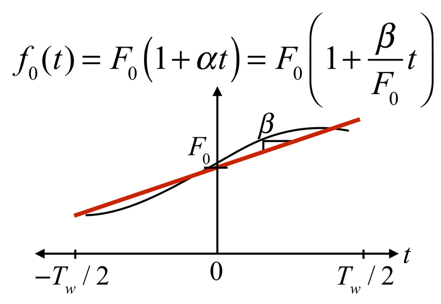


- Estimate complex time-frequency coefficients $\mathbf{\hat{X}_e(d,k)}$ of early reverberant component using Habets MMSE-LSA gains [1]:

$$\mathbf{\hat{X}_e(d,k) = G_{MMSE-LSA}(d,k) \cdot Y(d,k)} \qquad (2)$$

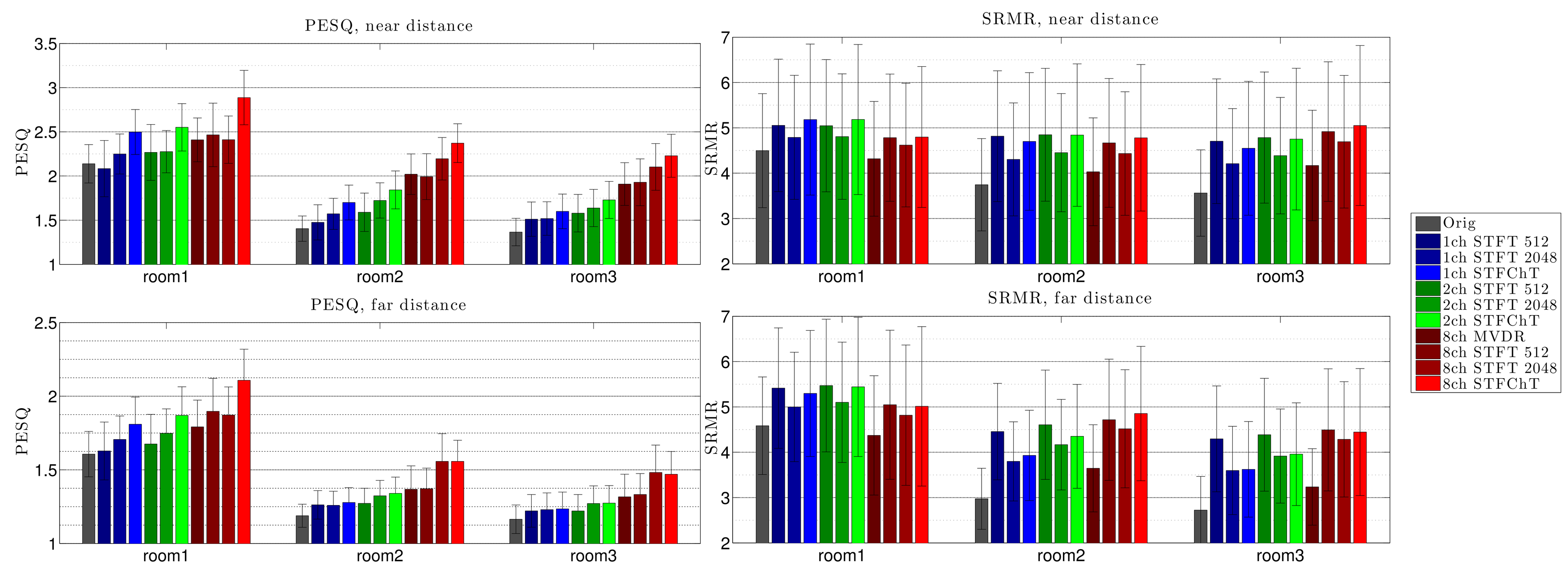
- Assumes stationary signal with constant $\mathbf{f_0(t)}$ over analysis duration.





- This assumption limits STFT window length, which limits data record length for statistical estimators.
- The STFChT is coherent with speech over longer durations, which allows longer data records and thus provides higher SNR.

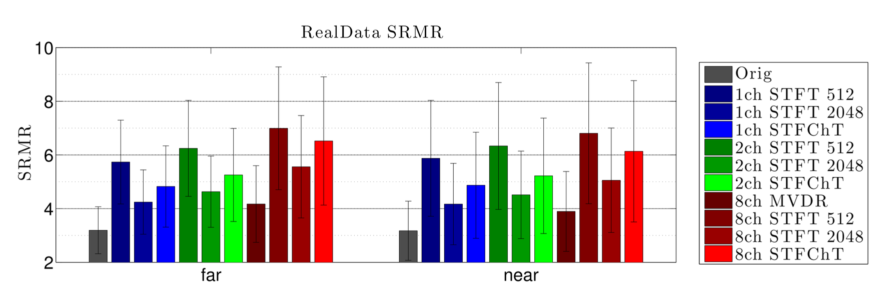
Results for SimData test set



STFChT processing achieves substantially better PESQ scores while maintaining roughly equivalent SRMR scores. STFT 512 uses a 32 ms window; STFT 2048 and STFChT use 128 ms windows.

Results for RealData test set

- STFT with standard 32ms window gives best SRMR scores, but introduces artifacts.



- STFChT processing, though achieving lower SRMR scores than STFT processing, exhibits fewer artifacts than STFT processing.

References

[1] E. A. P. Habets, "Speech dereverberation using statistical reverberation models," in *Speech Dereverberation*, Patrick A. Naylor and Nikolay D. Gaubitch, Eds. Springer, July 2010.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[3] H. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *Proc. IWAENC*, Tel Aviv, Israel, 2010, p. 1–4.

[4] M. Képesi and L. Weruaga, "Adaptive chirp-based time–frequency analysis of speech signals," *Speech Communication*, vol. 48, no. 5, pp. 474–492, May 2006.

This work was funded by U.S. Office of Naval Research (ONR) contract N00014-12-G-0078 delivery order 0013, and by U.S. Army Research Office (ARO) award W911NF-12-1-0277.