Spectral subtraction steered by multi-step forward linear prediction for single channel speech dereverberation

Keisuke Kinoshita, Tomohiro Nakatani and Masato Miyoshi
NTT Communication Science Laboratories

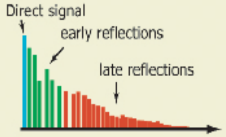
ABSTRACT

Feature of the proposed method:
- 1ch processing
- work only with a few seconds of training data
- work with low computational complexity

Result
- Substantial improvement in ASR performance in severe reverberant condition (RT_{20}=0.65 sec.)
- work robustly even in noisy environment

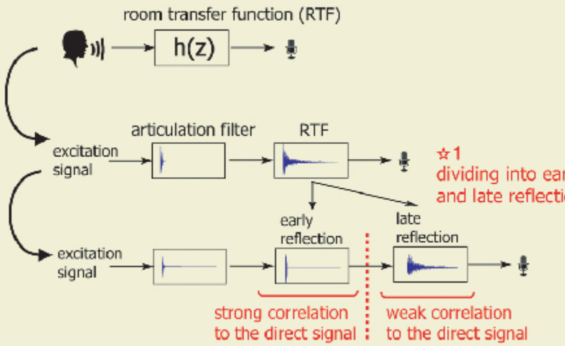
1. Problem of distant talk recognition

2. Which part is most harmful to ASR?



Early reflections: possible to handle (CMS, Delay-sum beamformer)
Late reflections: difficult to handle

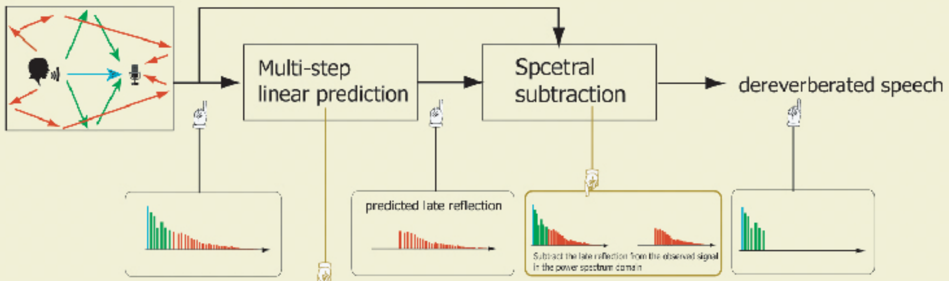
3. Idea of the dereverberation based on multi-step linear prediction

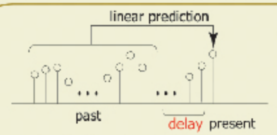


☆1 dividing into early reflections and late reflections

☆2: remove the late reflection based on the assumption that the direct signal and late reflections are uncorrelated

4. Overview of the proposed method





$$x(n)=\sum_{i=1}^{P}\alpha(i)x(n-i-D)+e(n)$$

x : observed speech α : prediction coefficent D : delay e : prediction error

calculating the proportion of the past signal within the present signal
=estimation of the late reflection

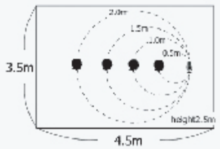
☺ The amplitude of late reflection can be estimated

☹ Theoretically, phase information can not be estimated
(see the paper about the details)

5. ASR experiment

5.1. Experimental conditions

5.1.1. reverberation condition
- simulation with image method

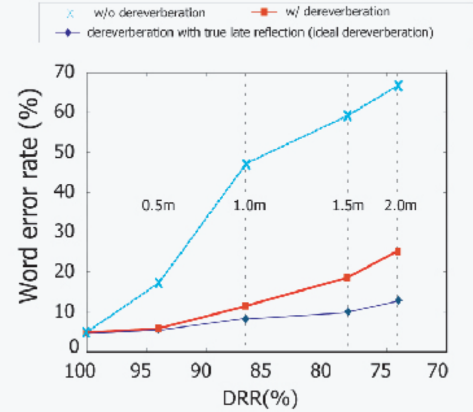


height of the microphone : 1.5m
height of the speaker : 1.5m
reverberation time : 0.65 sec.

5.1.2. dereverberation conditions
- sampling frequency: 12kHz
- estimate the filter with each sentence
- average sentence duration : 6 sec.
- filter length (AR coefficients) : 3000
- delay in linear prediction : 360 samples (=30ms)

5.1.3. ASR conditions
- large vocabulary continuous speech recognition
- Corpus:Japanese newpaper article speech (JNAS)
- Acoustic model :
  12th order MFCC and its energy, Δ, ΔΔ,
  3 state left-to-right HMM,3000 states,
  Gaussian mixture component: 16
- Language model : 20,000 words, standard trigram
- adaptation : Cepstral Mean Subtraction
- evaluation method : word error rate

5.2.Results



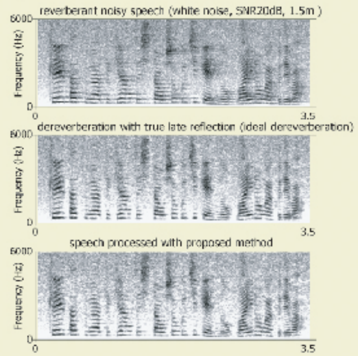
Direct to Reverberation Ratio (DRR)

$$DRR=\frac{\int_0^{25ms}h(t)^2dt}{\int_0^{\infty}h(t)^2dt}\times 100$$

h(t) : impulse response

6. Dereverberation in noisy environment

6.1 Improvement in spectrographic representation



6.2 Improvement in LPC cepstrum distance