"Making Machines Understand Us in Reverberant Rooms [Robustness against reverberation for automatic speech recognition]"

Yoshioka, T., Sehr A., Delcroix M., Kinoshita K., Maas R., Nakatani T., Kellermann W.

Signal Processing Magazine, IEEE, November 2012, pp.114-126

Andrea Zabaznoska

Advanced Signal Processing 2, Seminar
Institute of Signal Processing and Speech Communication
Graz University of Technology

June 11, 2013

Outline I

Outline II

INTRODUCTION – Necessity of Dereverberation

Distant-talking speech capturing

- ▶ Meeting speech recognition [techniques that allow the accurate automatic transcription and higher-level processing of multi-party meetings. Using prosody for extracting beyond-the-words information]

- ▶ Automatic annotation of videos [linguistic tagging or indexing]

- ▶ Speech-to-speech translation in teleconferencing

- ▶ Hands-free interfaces for controlling consumer products

- ▶ . . .

INTRODUCTION – Necessity of Dereverberation /cont'd

Automatic speech recognition ASR in distant-talking scenarios

- ▶ Background noise, competing speakers (additive and convolutional noise), microphone mismatch and room reverberation

- ▶ Problem *Reverberation cannot be captured by an additive or multiplicative term in the feature domain because reverberation has a dispersive effect on the speech feature sequences*

- ▶ In other words: *Reverberation spans a number of consecutive time frames and thus requires dedicated approaches*

INTRODUCTION – Methods for Combating Reverberation

Armin Sehr and Walter Kellermann

- ▶ Signal dereverberation and beam forming as preprocessing techniques

- ▶ Robust feature extraction and adjustment of the acoustic models to reverberation

- ▶ Reverberation modelling for speech recognition (combined approach)

- ▶ Audio processing and speech recognition, mainly driven by multidisciplinary approaches
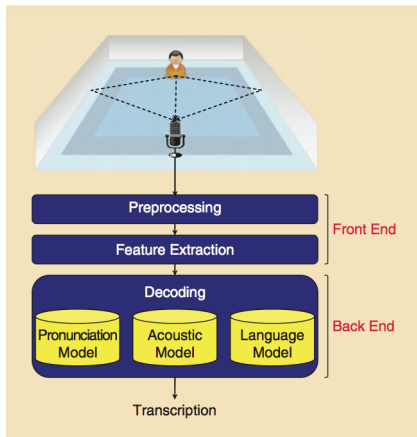
INTRODUCTION – Methods for Combating Reverberation /cont'd

Audio processing

- ▶ Blind deconvolution
- ▶ Nonnegative matrix factorisation

Speech recognition

- ▶ Noise robustness methods
- ▶ Novel ways for modelling reverberant data

Room Acoustics – Overview

Elements of room acoustics:

- ▶ Clean speech
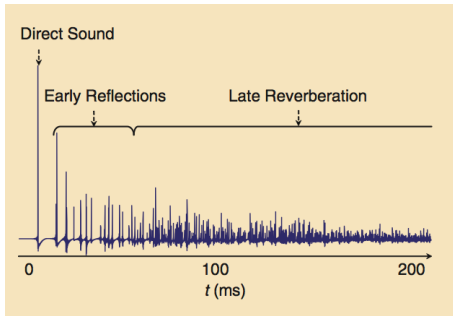- ▶ Room impulse response

$$y(t) = \sum_{\tau=0}^{T_h} h(\tau)x(t-\tau) + d(t) = h(t) \star x(t) + d(t)$$

- ▶ Reverberant speech signal
- ▶ Additive background noise

Blackboard . . .

Room acoustics – Overview /cont'd

Early reflections: h_i − Direct Sound

- Strong dependance on the speaker and microphone positions
- Occur within 50 ms after the direct sound

Late reverberation: h_l

- T_{60} ... reverberation time (from 200 to 1,000 ms)
- Decay exponentially and are independent of the positions

ROOM ACOUSTICS – Overview /cont'd

The insensitivity of the late reverberation magnitude to the speaker and microphone positions can be exploited to develop algorithms that are robust against speaker movement.

The energy ratio of the combined portion consisting of the direct sound and the early reflections to the late reverberation is measured by C_{50} and is highly correlated with speech recognition performance. This ratio mainly depends on the distance between the speaker and the microphone.
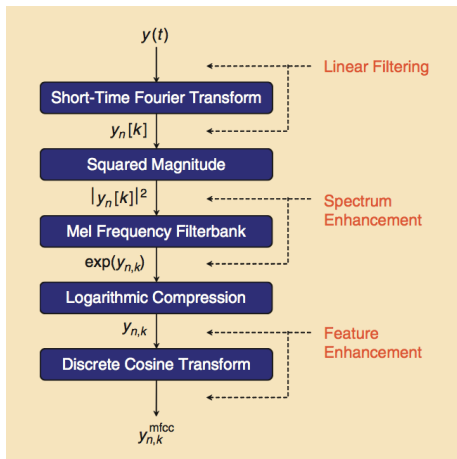
AUTOMATIC SPEECH RECOGNITION – Basic principles

▶ Extract the Mel-frequency cepstral coefficients (MFCCs):

$$y(t) \longrightarrow (\mathbf{y}_n^{mfcc})_n \in \mathbb{T}$$

▶ Transcribe the feature vector sequence by searching the sentence ω^* that maximizes the posterior sentence probability:

feat.vec.seq.   u.sequence   language model

$$\omega^* = \underset{\omega}{\mathrm{argmax}} \; p(\; (\mathbf{y}_n^{mfcc})_{n \in \mathbb{T}} \; | \; \omega \;) \quad p(\omega)$$

acoustic model

AUTOMATIC SPEECH RECOGNITION – Front-end

$y(t)$

Linear Filtering

Short-Time Fourier Transform

$y_n[k]$

Squared Magnitude

$|y_n[k]|^2$

Spectrum Enhancement

Mel Frequency Filterbank

$\exp(y_{n,k})$

Logarithmic Compression

$y_{n,k}$

Feature Enhancement

Discrete Cosine Transform

$y_{n,k}^{mfcc}$

MFCCs

1) Take the STFT of the signal

2) Emphasise the higher frequencies (increase the signal energy in the higher frequencies)

3) Mel Filter Bank Processing (please refer to next slide)

4) Take the DCT of the list of Mel log powers, as if it were a signal

AUTOMATIC SPEECH RECOGNITION – Front-end /cont'd

Mel Filter Bank Processing
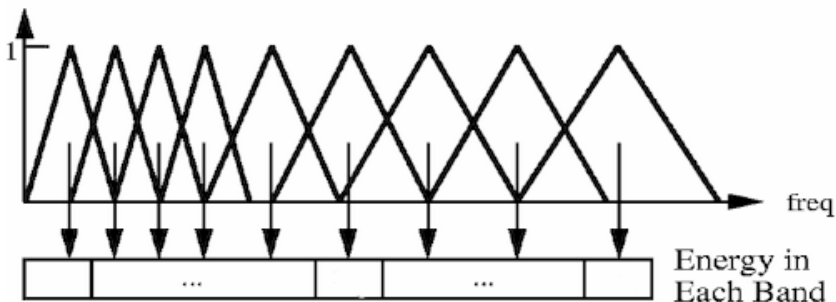
AUTOMATIC SPEECH RECOGNITION – Front-end /cont'd

Underlying sequence of discrete states, $(j_n)_{n \in \mathbb{T}}$:

$$p((\mathbf{y}_n^{mfcc})_{n \in \mathbb{T}}|\omega) = \sum_{(j_n)n \in T} \prod_{n \in \mathbb{T}} p(\mathbf{y}_n^{mfcc}|j_n)p(j_n|j_{n-1},\omega)$$

The conditional independence assumption means that the current feature vector depends only on the current state while the first-order Markov assumption means that the current state depends only on the previous state.

Reverberant speech recognition – Fundamental problem

Clean and reverberant log Mel-frequency filter bank features
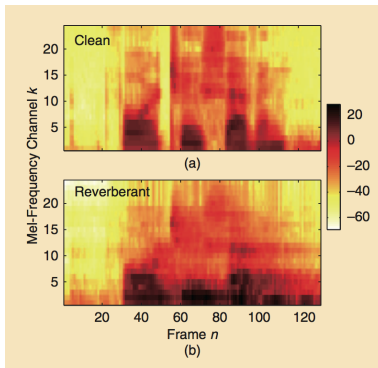
REVERBERANT SPEECH RECOGNITION – Fundamental problem /cont'd

Basic idea: recast :)

early reflections

$$y(t) = h_i(t) \star x(t) + h_l(t) \star x(t - \Delta) = h_i(t) \star x(t) + r(t)$$

late reverberation

late reverberation component

> Assumption: $r(t)$ is uncorrelated to $x(t)$ (additive noise). This approximation is partly justified by the fact that the autocorrelation coefficients of a clean speech signal are very small for time lags greater than 50 ms.

REVERBERANT SPEECH RECOGNITION – Fundamental problem /cont'd

Taxonomy of methods

Parallel model combination (PMC)

$\left.\vphantom{\begin{array}{c}a\\b\end{array}}\right\}$ extr.nonstat.

Vector Taylor series (VTS) compensation

Matched training

$\left.\vphantom{\begin{array}{c}a\\b\end{array}}\right\}$ acoustic context

Multistyle training

HMMs } conditional independence assumption

Reverberant speech recognition – Fundamental problem /cont'd

Extreme non-stationarity of the late reverberation $r(t)$

Renders almost all noise robustness techniques ineffective for computing late reverberation due to the fact that all techniques assume stationary or slowly varying noise to make noise parameter estimation and compensation possible

Account for the long-term acoustic content

Renders most of the training-based approaches as insufficient to successfully cope with reverberation.

Problem

Reverberant speech recognition – Fundamental problem /cont'd

## Account for the long-term acoustic content

Assumption: The mismatch between training and deployment environments can be reduced by the fact that it is expected similar or identical rooms to the target one to be included in the training environment

The conditional independence assumption between neighbouring feature vectors prevents the HMMs from effectively modelling the dependencies between reverberant feature vectors over several hundred milliseconds.

Solution: Use of dynamic features and extended feature vectors

REVERBERANT SPEECH RECOGNITION – Fundamental problem /cont'd

Extending the left context for each HMM state
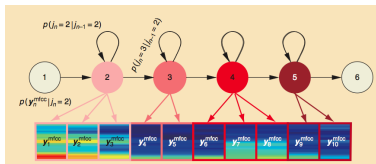
Increasing the left context by using tryphones is sadly not sufficient for this purpose $\Rightarrow$ use polyphones! But the number of polyphone models that would be needed to describe reverberant data would make a reliable training of the HMM parameters a computationally very challenging one, if not impossible.

Reverberant speech recognition – Fundamental problem /cont'd

Method comparison

- ▶ Problem stems from the assumption of conditional independence
- ▶ Late reverb is a very non-stationary additive interference and is predictable from past speech frames
- ▶ Long-term dependency is important
- ▶ Option 1: Remove the effect of late reverberation from observed feature vectors (long-term acoustic context) and Option 2: Change the acoustic model and decoder to deal directly with reverberant feature vectors

REVERBERANT SPEECH RECOGNITION – Improvement Measures

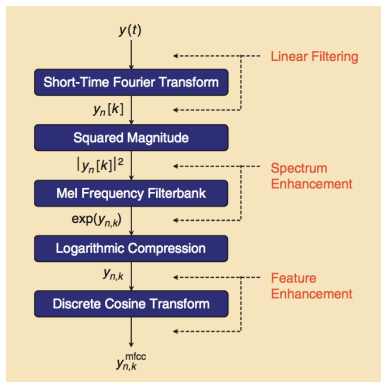| **Approaches**	**What to change**
Linear filtering	Waveforms or STFTs
Spectrum enhancement	Power spectra
Feature transformation	Feature vectors
HMM adaptation	HMM parameters
Context-aware decoding	Decoding algorithm

FRONT-END-BASED APPROACHES – Overview

Categorization

- *Linear filtering* ... dereverb. TD signals or STFT coeffs.

- *Spectrum enhancement* ... dereverb. corrupted power spectra (ignores signal phases)

- *Feature enhancement* ... removes reverb. directly from the corrupted feature vectors
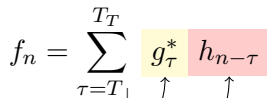
LINEAR FILTERING – Introduction

Basic approach

- ▶ Exploit both the amplitudes and phases of the signal (reverberation is a superposition of numerous time-shifted and attenuated versions of the clean signal)
- ▶ Exploit acoustical differences between multiple microphone positions

$y_n[k] \approx \sum_{\tau=0}^{T} h_\tau[k]^* x_{n-\tau}[k]$ ... the effects of reverberation may be represented as a one-dimensional convolution in each frequency bin, so the sequence $h_n[k]_{0 \leq n \leq T}$ can be viewed as an STFT-domain counterpart of the time-domain room impulse response.

LINEAR FILTERING – Blind deconvolution

Convolution of the room impulse response and the linear filter

$$f_n = \sum_{\tau=T_\perp}^{T_T} g_\tau^* \, h_{n-\tau}$$

filter coefficients

room impulse response

Objective: Set G so that f_n is nonzero if $n = 0$ and zero otherwise, while G is the set of adjustable filter coefficients.

LINEAR FILTERING – Blind deconvolution /cont'd

Long-term linear prediction (LTLP)

- ▶ Leverages a speech model that defines the pdf of a clean STFT coefficient x_n (normal distribution with zero mean and variance θ_n, where n is the frame index)

- ▶ Method of maximum likelihood $(G, \Theta = \theta_{n n \in \mathbb{T}})$ where $Y = (y_n)_{n \in \mathbb{T}}$ are the observed reverberant STFT coefficients

$$(\hat{G}, \hat{\Theta}) = \underset{(G, \Theta)}{\mathrm{argmax}} \log p(Y|G, \Theta)$$

LINEAR FILTERING – Blind deconvolution /cont'd

Long-term T_δ-step forward prediction (multistep prediction)

$$x_n = \sum_{\tau=T_\perp}^{T_T} g_\tau^* y_{n-\tau} \ldots \text{assume: } g_0 = 1 \text{ and } g_n = 0 \text{ for } T_\perp \leq n < T_\delta$$
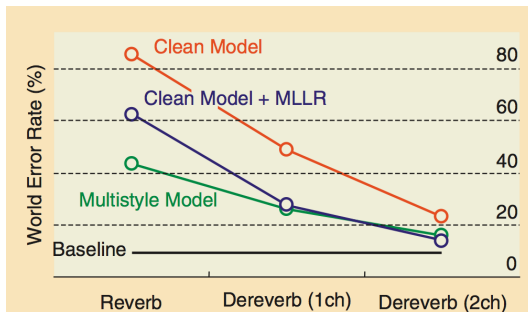
$$x_n = y_n + \sum_{\tau=T_\delta}^{T_T} g_\tau^* y_{n-\tau}$$

$$y_n = x_n + \sum_{\tau=T_\delta}^{T_T} g_\tau^* y_{n-\tau}$$

$$(\hat{G}, \hat{\Theta}) = \underset{(G,\Theta)}{\text{argmin}} \sum_{n\in\mathbb{T}} \left(\frac{\left| y_n - \sum_{\tau=T_\delta}^{T_T} g_\tau^* y_{n-\tau} \right|^2}{\theta_n} + log\theta_n \right)$$
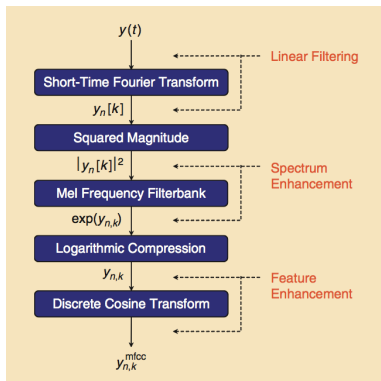
LINEAR FILTERING – Summary of LTLP

Method comparison

20K words (WSJ); $T_{60} = 0.78s$; Speaker-to-mic distance of 2 m;
No background noise

Spectrum Enhancement – Overview

- Basic idea: Given $(|y_n[k]|^2)_{n \in \mathbb{T}}$, restore the the clean power spectrum coefficients $(|x_n[k]|^2)_{n \in \mathbb{T}}$

- Late reverberations are insensitive to changes in speaker and microphone positions $\longrightarrow$ high robustness against speaker movement

- Can be combined with additive noise reduction techniques

Spectrum Enhancement – Overview /cont'd

Moving average estimator

Power spectrum reverberation model

$$y_n[k] \approx \sum_{\tau=0}^{T} h_\tau[k]^* x_{n-\tau}[k] \longrightarrow |y_n[k]|^2 \approx \sum_{\tau=0}^{T} |h_\tau[k]|^2 |x_{n-\tau}[k]|^2$$
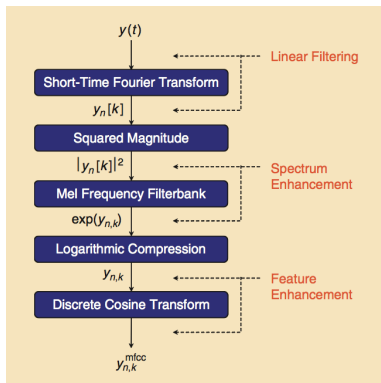
The power spectrum-domain representation of the room impulse response is known!

Predictive reverberation estimator

$$|r_n[k]|^2 = a[k] |y_{n-T_\delta}[k]|^2$$

T_{60} is known assuming a strict exponential decay of the late reverberation magnitude

FEATURE ENHANCEMENT – Overview

- ▶ Basic idea:
  Dereverberate features extracted from a reverberant signal. Based on a Bayesian framework, this technique attempts to infer the posteriori distribution of the clean features given the observation of all past corrupted features.

FEATURE ENHANCEMENT – Overview /cont'd

- ▶ The observation model relies on a simplified stochastic model of the RIR between the speaker and the microphone, having only two parameters (RIR energy and T_{60}, which can be estimated from the captured microphone signal)

- ▶ A hypothesis of a feature-domain model of reverberation is needed!
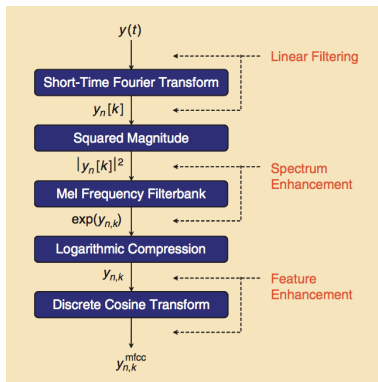
BACK-END-BASED APPROACHES – Overview

Categorization
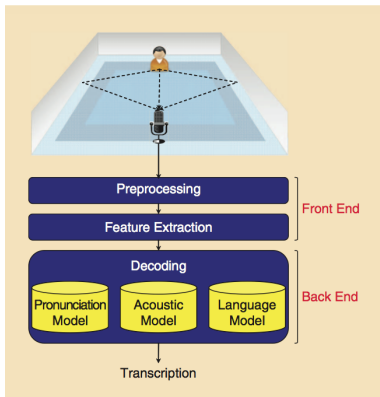
1. *HMM adaptation*: adjust the parameters of the acoustic model to the statistical properties of reverberant feature vectors $\longrightarrow$ use the adapted HMMs to transcribe reverberant utterances with a standard Viterbi decoder.

2. Tailor the decoder to the reverberant feature vector

BACK-END-BASED APPROACHES – Overview /cont'd

The functionality of the Viterbi decoder used in conventional speech recognisers relies on the evaluation of the emission pdf $p_Y(y_n|j)$, that is the likelihood of the observed feature vector y_n given state j.

It implies that the state likelihood is evaluated independently of preceding reverberant feature vectors $\longrightarrow$ HMMs cannot account for the long-term acoustic context inherent in reverberant feature vector sequences.

$p_Y(y_n|j, (y_\tau)_{\tau<n})$ is an emission pdf, in which the dependency on the feature vectors is explicitly stated.

Back-end-based Approaches – Overview /cont'd

Acoustic Context-Dependent Likelihood Evaluation

Frame-wise adaptation I

► Adjusts the means and the covariance matrices of the HMMs at each frame with regards to the preceding reverberant feature vectors

$$y_n \approx log(exp(\; h + x_n) + exp(\; r_n \;))$$

early reflection portion of the RIR
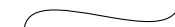late reverberation component of the reverberant speech

BACK-END-BASED APPROACHES – Overview /cont'd

Acoustic Context-Dependent Likelihood Evaluation

Frame-wise adaptation II

▶

predictor ... predicts the current late reverberation vector
from the previous observed feature vector

$$r_n \approx \boxed{a} + y_{n-1}$$

▶ h is a random variable with a normal distribution; a is
determined from the observed data $(y_n)_{n \in \mathbb{T}}$ using maximum
likelihood estimation

▶ $p_Y(y_n | j, (y_\tau)_{\tau < n})$ is modelled as a normal distribution with a
time-varying $\mu_{n,j}^Y$ and $\sum_{n,j}^Y$ at each frame

BACK-END-BASED APPROACHES – Overview /cont'd

Acoustic Context-Dependent Likelihood Evaluation

Frame-wise adaptation III

> The assumption of time-invariant reverberation model parameters is a drawback!
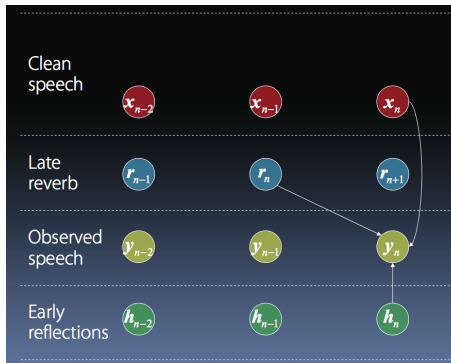>
> 1. The characteristics of the early reflections depend strongly on the speaker and microphone positions
> 2. Other time-varying errors in the approximation
>
> $\longrightarrow$ Use a time-varying model, i.e. statistical reverberation models, from which the model parameters (h and a) are sampled anew at each time frame
>
> $\longrightarrow$ Use REMOS!

REMOS – Introduction

A time-varying model

$\Rightarrow$ Use a time-varying definition to specify the emission pdf

$$p_Y(y_n|j, (y_\tau)_{\tau<n})$$
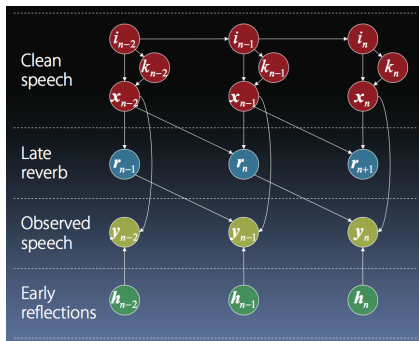
REMOS – Feature production model

A time-varying model
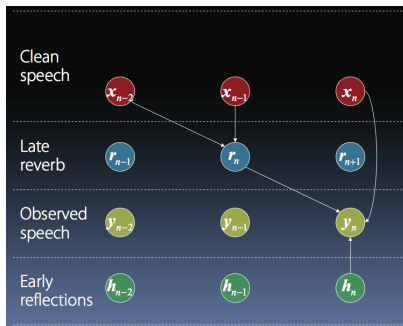
- $p_H(h), p_U(u)$
- $p_H(h), p_U(u)$ are learned from a set of RIRs or reverberant utterances
- Replace the predictive reverberation estimator with a moving-average reverberation estimator $r_n = log\left(\sum_{\tau=1}^{T} exp(\mu_\tau^H + x_{n-\tau})\right)$

$$log(exp(h_n + x_n) + exp(r_n + u_n))$$

REMOS – Feature production model /cont'd

A time-varying model

$$p_Y(y_n|j, (y_\tau)_{\tau<n}) =$$

$$\int p_\eta(y_n|x_n, (y_\tau)_{\tau<n}) \ p_\lambda(x_n|j)dx_n$$

effect of reverberation on the
clean vector x_n

emission pdf of the clean HMM

REMOS – Feature production model /cont'd

Marginalising all latent variables

$$p_Y(y_n|j,(y_\tau)_{\tau<n}) = \int p_\eta(y_n|x_n,(y_\tau)_{\tau<n})p_\lambda(x_n|j)dx_n$$

$$p_\eta(y_n|x_n,(y_\tau)_{\tau<n}) = \int \int p_H(h_n)p_U(u_n)$$
$$\cdot \delta(y_n - f_{mismatch}(x_n,(\hat{x}_\tau)_{\tau<n},h_n,u_n))dh_ndu_n$$

~~Marginalising~~ all latent variables $\Rightarrow$ *Estimating* all latent
variables!

REMOS – Feature production model /cont'd

Estimating all latent variables

$$p_Y(y_n|j, (y_\tau)_{\tau<n}) \approx p_H(\hat{h}_n) p_U(\hat{u}_n) p_\lambda(\hat{x}_n|j)$$

$$(\hat{h}_n, \hat{u}_n, \hat{x}_n) = \underset{(h_n, u_n, x_n)}{\text{argmax}}\ p_H(h_n) p_U(u_n) p_\lambda(x_n|j)$$

$$\text{subject to } y_n = f(x_n, (\hat{x}_\tau)_{\tau<n}, h_n, u_n)$$

For each frame n and each state j, the Viterbi score is calculated by first solving the optimisation problem and then evaluating the emission pdf.

y_n is decomposed into contributions x_n from the clean HMM, the feature-domain of initial reflections h_n and r_n depending on state j.

REMOS – Conclusions

Different clean feature estimates for different states by leveraging the statistical model of reverberation parameters
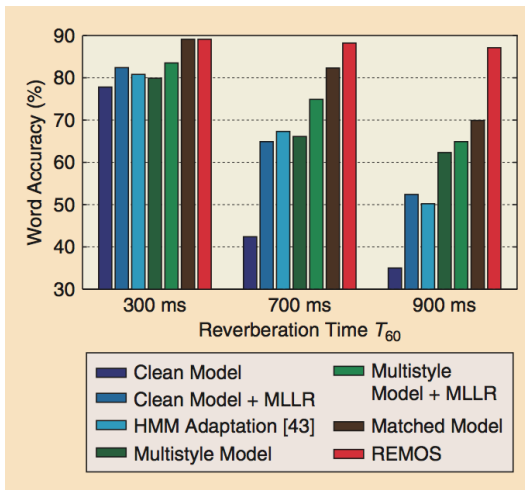
👍 High word accuracies even in severely reverberant environments

👍 High flexibility to changes of speaker positions and room changes

👍 Possible expansion by adding a noise model to the REMOS framework

👎 Requires a recogniser trained on a static log Mel-frequency filter bank features and single Gaussian densities

REMOS – Conclusions /cont'd

Summary & Conclusio – I

Pros & Cons : front-end

👍 Making changes @front-end doesn't require any modifications @back-end

👍 Computational complexity is independent of the acoustic model size

👍 Easily combined with advanced recognition techniques such as fMPE

👎 Estimation errors degrade the decoder's performance

Pros & Cons : back-end

👍 Less prone to estimation errors

👍 Coherent performance of both speech recognition and reverberation compensation

👎 Computational complexity proportional to acoustic model parameters

👎 Conventional features make them hardly to combine with advanced techniques

Summary & Conclusio – II

✓ Reverberation can be modelled as additive interference

✓ The main difference from common noise and interference is its extreme non-stationarity $\longrightarrow$ fundamental problem

✓ Long-term dependency is important: strong relationship between long-term consecutive reverberant frames is an essential clue to compensate for reverberation

Future Research Topics

- ▶ Reverberant speech recognition as subproblem of transcribing distant-talking speech

- ▶ Combination of different approaches & extension to state-of-the-art-systems that use discriminative or posterior-based features such as fMPE (Minimum Phone Error)

- ▶ Joint compensation of additive noise and reverberation

References I

P.A. Naylor, N.G. Gaubitch,
"Speech Dereverberation,"
Berlin: Springer-Verlag, 2010.

T. Nishiura, Y. Hirano, Y. Denda, M. Nakayama
"Investigations into early and late reflections distant-talking speech recognition toward suitable reverberation criteria,"
Proc. Interspeech, pp. 1082–1085, 2007.

H. Kuttruff
"Room Acoustics,"
5th ed., Abingdon, Oxon: Spon Press, 2009.

T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, A. Nakamura, J. Yamato
"Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera,"
IEEE Trans. Audio, Speech, Language Process., vol.20, no.2, pp. 493-513, 2012.

B.E.D. Kingsbury
"Robust speech recognition using the modulation spectrogram,"
Speech Commun., vol.20, no.1-3, pp. 117-132, 1998.

References II

A. Acero, L. Deng, T. Kristjansson. J. Zhang
"HMM adaptation using vector Taylor series for noisy speech recognition,"
Proc. Int. Conf. Spoken Language Process., pp. 869-872, 2000.

M.J.F. Gales, S.J. Young
"Robust continuous speech recognition using parallel model combination,"
IEEE Trans. Speech Audio Process., vol.4, no.5, pp. 352-359, 1996.

C. Legetter, P. Woodland
"Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,"
Comput.Speech Language, vol.9, no.2, pp. 171-185, 1995.

J. Gauvain, C-H. Lee
"Maximum aposteriori estimation for multivariate Gaussian mixture observation of Markov chains,"
IEEE Trans. Speech Audio Process., vol.2, no.2, pp. 291-298, 1994.