Compensation for reverberation by human and machine listeners with synthetic and naturalistic speech

Aalto University · Finland · 17 April 2012

Amy Beeston and Guy Brown

This work is licenced under the Creative Commons Attribution-NonCommercial-NoDerivs 2.0 UK: England & Wales License. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/2.0/uk/
collaborators

- University of Sheffield
  Speech and Hearing
- University of Reading
  Anthony Watkins, Simon Makin, Andrew Raimond
- Advisors
  Kalle Palomäki, Hynek Hermansky, Phonak
- EPSRC funding
1. Compensation for reverberation: Watkins’ sir-stir paradigm
2. Modelling sir-stir
3. Generalising to naturalistic speech
4. Reverberation robust front-end for ASR
Compensation

1. Compensation for reverberation: Watkins’ sir-stir paradigm
2. Modelling sir-stir
3. Generalising to naturalistic speech
4. Reverberation robust front-end for ASR
Compensation for reverberation

introduction

• reverberation degrades speech intelligibility
  - acoustic content differs with distance
  - but phonetic content persists

• we can compensate for reverberation

• compensation is reliant on contextual sound
  - what factors promote/inhibit compensation?

• can machine listeners use equivalent cues?
  - same mistakes as humans?
in reverberation

stop consonants

- sensitive to reverberation
- identification depends on rapid amplitude modulation (envelope)
- overlap- and self-masking in reverb
  - peaks are prolonged
  - dips are filled

Watkins’ approach

- real room reverberation
- synthetic speech material

Watkins’ stimuli

• synthetic test word continuum
  - spoken ‘sir’ has /t/ artificially imposed for ‘stir’

• sir-stir identification depends on envelope: rapidly changing amplitude modulation (AM)
• reverberation reduces AM
• test word embedded in predictable location
  - one recorded context sentence
• test and context reverberation varied unpredictably

sir-stir

experimental paradigm

test word continuum

into context sentence

category boundary

“stir”

“sir”
sir-stir

test word reverberated

near context, far test

“stir”

“sir”
context also reverberated
compensation effect

increased reverb on test word

increased reverb on context

compensation

far test

near test
Modelling sir-stir

1. Compensation for reverberation: Watkins’ sir-stir paradigm
2. Modelling sir-stir
3. Generalising to naturalistic speech
4. Reverberation robust front-end for ASR
modelling sir-stir

reverberation and dynamic range

- reverberation fills dips in temporal envelope
- compensation as a restoration of dynamic range?
- decr. dynamic range => incr. mean to peak ratio (MPR)

near condition
- little reverberation
- large dynamic range
- low MPR value

far condition
- more reverberation
- small dynamic range
- high MPR value
modelling sir-stir

possible role of efferent system

- efferent system implemented in control of dynamic range
- compensation as restoration of dynamic range?

MPR of simulated auditory nerve response $\propto$ reverberation
summed across bands (frequency independent)

\textbf{Ferry and Meddis (2007).} J Acoust Soc Am, 122(6), 3519-3526.
\textbf{Beeston and Brown (2010).} Interspeech, pp 2462-2465.
modelling sir-stir

efferent feedback loop

- incr. reverb => incr. MPR => incr. ATT => decr. output from DRNL => rate curve shifts to right
- => response to stimuli with low sound level falls below threshold
modelling sir-stir
recovering dips

- helps to recover the /t/
• **Reverse speech**: compensation persists
  MPR prior to test word is unaltered by speech direction
**Reverse reverberation:** compensation is abolished

MPR prior to test word is reduced in this case

![Graph showing category boundary vs context distance for forward and reverse speech with forward and reverse reverberation, with data points for human near test, human far test, model near test, and model far test.]
within-channel model

frequency importance i.

- new data from Watkins’ lab using 8-band stimuli
- reverberation applied only on certain bands
- high freq. bands most important in sir-stir distinction

within-channel model

frequency importance ii.

- frequency sensitivity of efferent system
- physiological data *(for a cat, not a human)*
  - approx. linear increase in region 100-8000 Hz

Within-band (frequency dependent) model
- MPR typically incr. with reverb in every channel
- ATT determined channel-by-channel
modelling sir-stir

interim conclusions

• compensation is not reliant on phonetic information
• compensation effects might be modelled with a system that monitors dynamic range
• Watkins’ synthetic stimuli
  - same context sentence and talker
  - artificially created /t/ in ‘stir’
• do Watkins’ findings generalise?
Naturalistic speech

1. Compensation for reverberation: Watkins’ sir-stir paradigm
2. Modelling sir-stir
3. Generalising to naturalistic speech
4. Reverberation robust front-end for ASR
• do Watkins’ findings hold for naturalistic speech?

• Articulation Index (AI) Corpus
  - includes sir and stir
  - more context words
  - more talkers

• each AI corpus utterance uses different talker, vocabulary, speech rate, pitch contour, stress pattern etc.
  - cancel excess variability?
  - analyze results with regard to this variability?

naturalistic speech

ideals

• naturalistic speech
  - real world listening
  - ASR compatible

• increase data per participant
  - increase subset of Articulation Index Corpus
  - further consonant/vowel sets?

• minimize manual handling
  - word boundaries located via (HTK) forced-alignment
naturalistic speech

extending sir-stir

• subset of corpus
  sir · skur · spur · stir

• consonants
  unvoiced stops, diff. place of articulation
  /p/ front · /k/ back · /t/ middle

• vowels
  with {s, sk, sp, st} can have {R, e, i, E, I, @, (a, o)}
naturalistic speech

relative information transmitted (RIT)

- no category boundary
- misclassifications

<table>
<thead>
<tr>
<th></th>
<th>@ nf</th>
<th>sir</th>
<th>skur</th>
<th>spur</th>
<th>stir</th>
</tr>
</thead>
<tbody>
<tr>
<td>sir</td>
<td>37</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>skur</td>
<td>6</td>
<td>29</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>spur</td>
<td>16</td>
<td>3</td>
<td>19</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>stir</td>
<td>16</td>
<td>2</td>
<td>1</td>
<td>21</td>
<td></td>
</tr>
</tbody>
</table>

- RIT
  - regards participants as channels
    - accept input stimuli
    - produce output responses
  - measures their information transfer characteristics

naturalistic speech

experiment 1: ‘cutoff’

- Possible to replicate compensation for reverb?
- Necessary to increase overall error rate?
  low pass filtered to avoid ceiling effects
- same and mixed distance sentences
  {near, far} context + {near, far} test
  {1, 1.5, 2, 3, 4} kHz low-pass filter cutoff
- 1600 stimuli partitioned across 20 listeners (N=40)
  4 targets X 20 talkers X 4 distances X 5 filters
naturalistic speech

experiment 1: ‘cutoff’

- errors incr. as low-pass filter cutoff frequency decr.
- compensation apparent when high freqs are present

![Graph showing error metric for different frequencies and distances.](image)

N=40

1 kHz 1.5 kHz 2 kHz 3 kHz 4 kHz

error metric: 1-RIT

- test 10 m
- test .32 m
naturalistic speech

word-level analysis: ‘cutoff'

- Allen and Li: \{/t/, /k/, /p/\} identified by burst frequency
  /t/ at 4 kHz; /k/ at 1.4 – 2 kHz; /p/ at 0.7 – 1kHz
- /k/ had generally fewer errors (but advantage was lost at low freqs)
- /p/ holds identity better at 1.5 kHz
naturalistic speech

experiment 2: ‘reverse’

- do time-reversal procedures disrupt compensation if applied to preceding context?

- time reversed speech and/or reverberation
  - fwd reverb: CW overlaps TEST
  - rev reverb: no overlap

Watkins’ condition

- 1280 stimuli partitioned across 16 listeners (N=64)
  4 targets X 20 talkers X 4 distances X 4 reversals
• compensation is present for forward reverberation, but abolished with reverse reverb?
naturalistic speech

experiment 3: ‘inAndExtrinsic’

• contribution of intrinsic (test), extrinsic (context) info?
  – following context (ex): removed
  – preceding context (ex): near, far or silenced compensation forwards in time?
  – test word (in): reverb tail either full or gated gate cuts tail at end of test-word’s vowel retroactive effects?

  ![Add tail >](near)
  ![Cut tail >](far)
  ![Gated](gated)

• 5760 stimuli partitioned across 12 listeners (N=48)
  {near, far, silent} context X {near, far} test
  4 consonants X 6 vowels X 20 talkers X 3 contexts X 2 tests

17 Apr 2012 · Aalto University · Finland · Amy Beeston
naturalistic speech

experiment 3: ‘inAndExtrinsic’

- compensation occurs without following or preceding context (‘silent’ acts like ‘far’)
- Intrinsic info used when context is ambiguous, e.g. missing or inappropriate

<more>

17 Apr 2012 · Aalto University · Finland · Amy Beeston
interim conclusions

- compensation for reverberation exists for naturalistic speech despite -
- high variability (cf. Watkins)
  - more talkers
  - more context words
  - more test words
- different things going on for different test words
- analysis methods require still more thought!
1. Compensation for reverberation: Watkins’ sir-stir paradigm
2. Modelling sir-stir
3. Generalising to naturalistic speech
4. Reverberation robust front-end for ASR
• use auditory model as a ‘compensation’ front-end
  - see if recognition of stop consonants improves
reverberation robust front-end

HSR/ASR

- Compare human/machine performance
  - assess consonant confusions on same listening task
  - experiment 1: ‘cutoff’ 4kHz condition (N=20)

- Baseline system <details>
  - HTK/MFCC phone recogniser

- Simplified auditory model <details>
  - efferent circuit engaged (with more ATT at far distance)

---

reverberation robust front-end

ASR baseline system

- human: compensation
- machine: does not

ASR error \( \propto \) amount of reverb on test word
near-near < far-near < near-far < far-far

\[1 \rightarrow \text{information transmitted} \]

\[0 \rightarrow \text{context distance (m)}\]

\[50 \rightarrow \text{percentage error} \]

\[0 \rightarrow \text{context distance (m)}\]

\[0 \rightarrow \text{1-relative information transmitted} \]

\[0 \rightarrow \text{context distance (m)}\]

\[\bullet \text{human test 0.32m} \]
\[\bullet \text{human test 10m} \]
\[\bullet \text{MFCC test 0.32} \]
\[\bullet \text{MFCC test 10m} \]

reverberation robust front-end

simplified auditory model

- no efferent system: auditory features are similar to MFCC result, no compensation

- with efferent system, 4 dB ATT at far context: a little compensation if viewed with %ERR (but not with RIT)

reverberation robust front-end
parallel-likelihood model

• compensation: process by which listeners dynamically weight evidence from different speech models?
• simulation of compensation
  - tracking dynamic range of context’s temporal envelope
  - via model selection
• performance is good when context and test word conditions are matched.
• but a mismatch (at near-far) => incorrect acoustic model => more consonant confusions.

reverberation robust front-end
parallel-likelihood model

Signal -> MFCC

Likelihood from near model -> $p(x(t) | \lambda_n)$

Detect reverb conditions

Likelihood from far model -> $p(x(t) | \lambda_f)$

Likelihood weighting

Phone sequence -> Decode
Table 1: Confusion matrices for human listeners (left), computer model given oracle information about the reverberation condition (center) and computer model that uses a MPR metric to determine the reverberation condition (right). Reverberation conditions are labelled as context-test distance. Rows correspond to the stimuli presented; columns record the responses.
human listeners use information from preceding context to effect compensation

conventional ASR systems do not

simulation of compensation
- tracking dynamic range of context
- perhaps a number of auditory processes involved

much work remains
- frequency-dependent effects
- wider range of consonant confusions
the end

thank you for listening


abstract

Compensation for reverberation by human and machine listeners with synthetic and naturalistic speech material

Humans exhibit a form of auditory constancy in reverberant rooms, deriving information from contextual sounds prior to a spoken test word [Watkins (2005) J Acoust Soc Am 118, 249-262]. Watkins has demonstrated the effect of compensation for reverberation in a series of experiments where a synthetic test word is identified either as 'sir' or 'stir' depending on clues gleaned from the preceding speech.

Inspired by principles of human audition, a computational model is described which displays human-like compensation patterns for Watkins' stimuli with forward and time-reversed speech and/or room-reflections. In conjunction with a simple template-based speech recogniser, the auditory model uses efferent suppression to reduce the influence of room reverberation [Ferry & Meddis (2007) J Acoust Soc Am 122, 3519-3526].

To approach the real-world conditions required for automatic speech recognition (ASR) studies, recent perceptual experiments carried out in Sheffield have investigated compensation for reverberation using naturalistic speech with more talkers and with a larger vocabulary for both context and test words. Finding evidence of compensation in the pattern of consonant confusions that listeners make, ongoing work to model this data is now aimed at developing a 'constancy' front end for reverberant ASR.