Missing Links between Hearing and Robust Speech Recognition

Ning Ma (in collaboration with Jon Barker and Phil Green)

Speech and Hearing Research Group (SPandH)
Department of Computer Science
University of Sheffield, UK
Speech Recognition in Daily Life

In everyday listening conditions speech is naturally combined with various noises and reverberation. Human listeners are remarkable adept at recognising speech in such conditions.

Imagine yourself being in a party... what can you hear?

How do we recognise what one person is saying when others are speaking at the same time (the “cocktail party problem”)? On what logical basis could one design a machine (“filter”) for carrying out such an operation? – Cherry (1953)
A major problem with speech recognition technology is robustness, i.e. ability to sustain accuracy in adverse conditions, as human listeners do. In SPandH we have been developing an approach to this problem which is inspired by computational models of human hearing. This talk tells the story of this research.

The outline of this talk is:

- Robustness issues with speech recognition
- Computational models of hearing
- ‘Missing data’ methods for robust speech recognition
- The ‘speech fragment decoding’ framework
Automatic Speech Recognition

- The process of converting spoken words to machine-readable input

  \[ \hat{W} = \arg \max_W P(W|X) \]

  \[ = \arg \max_W \frac{P(X|W)P(W)}{P(X)} \]

  \[ = \arg \max_W P(X|W)P(W) \]

- The standard statistical ASR framework
  - Convert audio waveform into a sequence of acoustic observations
    \[ X = x_1, x_2, x_3, \ldots, x_T \]
  - The goal is to find the most probably word sequence

Bayes’ rule

Acoustic model

Language model
Automatic Speech Recognition (2)

- Most successful ASR systems employ Hidden Markov Models (HMMs)
  - A sequence of stationary states $Q = q_1, q_2, \ldots, q_N$
  - Output observation distributions and transition probabilities

- HMMs are simplified by two assumptions
  - Each state is conditionally independent of all other states given the previous state (the first-order Markov assumption)
  - Each observation is conditionally independent of all other observations given that state which generates it

$$\hat{Q} = \arg \max_Q P(X|Q) P(Q) = \arg \max_Q \prod_{i=1}^{T} P(x_i|q_i) \prod_{i=2}^{T} P(q_i|q_{i-1}) P(q_1)$$
ASR performance can be excellent in friendly conditions. However, there are many ASR applications where these conditions don’t apply:

- Using an ASR device in a noisy environment (e.g. an airport)
- Dealing with informal speech, uttered to be recognised by humans rather than specifically for machine recognition (e.g. a phone call)
- Recording from “far-field” microphones (e.g. in a meeting)
- Speech in reverberant environments (e.g. lecture rooms)
• Only a small amount of noise (signal-to-noise ratio 10 dB) is needed to blow away a conventional recogniser

• Lack of robustness is perhaps the biggest single factor limiting commercial take-up of ASR.
Mismatch between Training & Testing

The conventional way of thinking about the robustness problem is:

“There is a mismatch between training and testing conditions. What we have to do is find a way of eliminating, or at least reducing, the mismatch.”

- Noise-robust features: signal processing which is less affected by noise
- Speech enhancement: ‘clean up’ the speech before you recognise it
- Model compensation: change the models to reflect the noise conditions.
Noise-resistant Acoustic Features

- Look for signal processing techniques which are insensitive to changes in the environment, i.e. make the mismatch smaller

- Example: RASTA (RelAtive SpecTrA) processing
  - Based on the fact that changes in the speech envelope (modulations) lie in a characteristic range (4-50Hz) – the syllabic rate
  - Any modulation component outside that range can not be due to speech
RASTA Filtering

- System essentially deaf to near-stationary noises and to rapidly-varying noise (band-pass filtering)
- Assumes noise modulation outside speech range
- Not true for many noise sources, e.g. another speaker

Hermansky (1994)
Speech Enhancement

- Try to ‘clean up’ the noisy speech, making it more like clean speech
- Spectral Subtraction is the most popular technique of this kind
  - Obtain an estimate of the noise power spectrum
  - Estimate the clean speech power spectrum by subtracting the noise means in each frequency channel in each frame
- Recognise using models trained on the clean speech spectrum
Spectral Subtraction

- **Assumptions**
  - The noise power spectrum is available: if it is steady or slowly changing it can be measured from periods in which there is no speech
  - Linearity – hence the subtraction makes sense

- **Problems**
  - Assumes that noise has zero variance – always subtract mean
  - Sometimes this is wrong – leads to ‘musical noise’
  - Not good for non-stationary noise
Model Compensation

- Reduce the mismatch by using models that accommodate the noise
- Multicondition training
  - Train on clean speech with a variety of typical noises added
  - Can improve robustness, maybe at the cost of reducing clean speech performance
- HMM Decomposition and Parallel Model Combination
Parallel Model Combination

- Obtain a statistical model of the noise
- Combine clean speech models with noise models to resemble those you would have obtained if you had trained on speech plus noise
- Can produce very good results if assumptions are met
- Problems
  - Assumes sufficient is known about the noise to make an adequate model
  - Hard to apply if noise is unpredictable
  - Factorial nature of model combination – the combined state space expands exponentially
Auditory Scene Analysis

- In contrast to ASR, human listeners can recognise speech accurately in very noisy conditions (the ‘cocktail party’ problem)

- Humans (and many animals) are adept at *Auditory Scene Analysis*. At any time, the listener is aware of
  - What sounds are present in her/his auditory surroundings
  - The nature of these sounds (it’s a voice, it’s the wind, it’s a car...)
  - The sound source direction
  - How the sounds are changing (moving around, getting fainter...)

- Analogy to vision

A. Bregman (1990)
Auditory Scene Analysis (2)

- The result of ASA is to organise the scene into streams, each corresponding to a set of time-frequency regions which correspond to a single sound source – auditory grouping
- The listener can selectively attend to one stream or another, even to faint ones
- ASA is highly developed – it provides an evolutionary advantage
  - The auditory system routinely solved the ASA problem, in real time
Auditory Scene Analysis (3)

• Listeners solve the ASA problem by interactively exploiting
  • Primitive data-driven grouping principles – innate constraints driven by various properties of the acoustic input
    Bottom-up cues: pitch, common fate, direction
  • Schema-driven constraints – prior knowledge of familiar patterns that have been learnt from acoustic environments
    Top-down cues: language, the African language involving clicks

• Can ASA be the key to robust ASR?
  • We need to say a little about how hearing works
Overview of the Auditory System

- The outer ear consists of the external ear (pinna), ear canal and eardrum
- The pinna collects sound and directs it into the ear canal
- Captured sound causes the eardrum to vibrate
- The vibrations are transmitted to the cochlea by the 3 bones (ossicles) of the middle ear
The Cochlea

- The cochlea in the inner ear is the ‘organ of hearing’
  - Filled with fluid and divided along its length by the basilar membrane
  - Incoming sound produces waves in the cochlea fluid which cause displacement of the basilar membrane, and position of maximum displacement in the cochlea encodes sound frequency

- Hair cells transduce basilar membrane motion to neural firing in the auditory nerve
Auditory Filterbank

- This highly redundant representation in the auditory nerve is in complete contrast to orthogonal representations like MFCCs.

- The characteristics of auditory frequency analysis have been studied by psychoacoustics and neuroscience.

- They can be modelled by gammatone filters to produce a cochleagram – the auditory equivalent to a spectrogram.
The auditory scene is somewhat like a visual image
  - With time on the x-axis and frequency on the y-axis
  - A ‘pixel’ in this image indicates how much energy there is in a particular frequency band at a particular time

In the cochleagram frequency resolution is much greater below ~1k Hz than above

The first few harmonics are resolved
The goal of CASA is to use computational models to understand how the auditory system performs ASA (Cooke, 1991; Brown, 1994).

- Primitive grouping vs. schema-driven grouping
- A primary cue for primitive grouping is harmonicity
  - If a sound is produced by a periodic source, the energy will be concentrated at the fundamental frequency and the harmonics.
Linking ASA with Robust ASR

- Can ASA be used as the basis for a robust ASR technique?

- Attractions:
  - Primitive CASA uses only low-level constraints (harmonicity, source direction) which are derived from physics and the properties of the auditory system. It does not require models for the noise.
  - Apply equally well to all sounds
  - There’s no need to assume how many sound sources there are
The Missing Data Problem

- CASA will never be able to recover all the speech from a mixture of speech and noise
  - Some speech evidence is effectively masked by noise
  - Here’s a speech + siren cochleagram

- But is ‘missing data’ an issue to robust ASR?
Masking in Auditory Mixture

- Speech energy is sparsely distributed in the time/frequency plane
  - Concentrates on local regions such as harmonics and formants
  - Overlapping in time becomes less a problem in the T/F plane

- Speech encoding is redundant the auditory system
  - Speech remains intelligible even when a large part of the speech spectrum is removed (Fletcher, 1953; Warren, 1995, 2000)
  - The redundancy allows listeners to perceive speech in noise based on relatively sparse information
The Disjoint Allocation Assumption

- Unlike visual signals, sounds are additive rather than opaque
- Any time-frequency ‘pixel’ is dominated by a single source
- Justification:
  - Speech energy is concentrated in sparse regions
  - The compressive processing used in representations like the cochleagram: the log-max approximation $\log(a+b) \approx \max(\log(a), \log(b))$
  - The situation resembles visual occlusion

![Image of cochleagram](image)

Granada 2010
The problem is to identify those ‘reliable’ pixels – those dominated by the speech source: the result is called the ‘missing data mask’

Can use pre-mixed signals to find the ‘oracle mask’ – cheating!

Can use CASA and noise estimates (if available) to obtain missing data masks without cheating

- Local SNR estimation
- Harmonicity
- Common onsets/offsets
- Spatial locations
Adapting HMM-based ASR for MD

• If some of the features in the observation are missing (i.e. unreliable), \( x \) can be partitioned into reliable & unreliable components:
  \[
  x = (x_r, x_u)
  \]

• Data Imputation
  - Try to reconstruct the missing data \( x_u \), then recognise as normal
  - Advantages:
    After imputation, can convert from spectral to cepstral domain
    Can make use of an existing, pre-trained recogniser.

• Marginalisation
  - Estimate state likelihood directly, using the reliable values and constraints on the missing values
  - Advantage:
    Can retain a probabilistic framework rather than settling for a single value
Marginalisation-based Techniques

- Classification is based on the marginal distribution of the reliable features by integrating over the unreliable components $x_u$:

$$p(x|q) = \int p(x|q) dx_u = \int p(x_r, x_u|q) dx_u$$

- If the state distribution $p(x|q)$ is modelled by Gaussian mixtures

$$p(x|q) = \sum_{k=1}^{M} P(k|q) \prod_{i \in r} p(x_i|q, k) \prod_{i \in u} \frac{1}{x_i} \int_{0}^{x_i} p(\hat{x}|q, k) d\hat{x}$$

\[
\begin{aligned}
\int_{0}^{x_i} p(\hat{x}|q, k) d\hat{x} \\
p(x_i|q)
\end{aligned}
\]
Counter-evidence from Bounds

- The bounds integral can be thought of as a measure of counter-evidence
  - State $q$ matches the reliable evidence well but there is insufficient energy in the unreliable components
Soft Missing Data Masks

- The binary reliable/unreliable decision for a pixel might be wrong
- Replace it by an estimate of the **probability that it’s a speech pixel**
- Use this probability to weight the evidence and counterevidence

\[
p(x|q) = \sum_{k=1}^{M} P(k|q) \prod_{i=1}^{N} \left( w_ip(x_i|q, k) + (1 - w_i) \frac{1}{x_i} \int_{0}^{x_i} p(x'|q, k)dx' \right)
\]

- Gives a significant win (Barker, 2001)
• ‘Missing Data’ is a common problem in Science and Engineering
  • Unreliable sensors, band-restricted transmission, incomplete surveys
Some Challenges with MD ASR

- Requires complete speech/noise segregation
  - Mask estimation is a challenging problem
  - Noise can be highly unpredictable and variable
- Segregation and recognition are decoupled
  - In contrast to human speech recognition
  - Prior knowledge in top-down models is ignored when the segregation hypothesis is formed

**Mask estimation is challenging!**

**Only bottom-up processing**
A Visual Example of MD ASR

- Recognising “SPEECH” from noisy data
  - Need to identify a correct MD mask (bottom-up grouping)
Human speech recognition is in contrast

- Interactively governed by bottom-up process that exploit common characteristics of the acoustics and top-down constraints employing the knowledge of familiar patterns

- The ability of Auditory Scene Analysis

Barker et al. (2005) proposed the Speech Fragment Decoding framework

- Combines bottom-up & top-down processing
- Source separation & recognition are tightly coupled
Speech Fragment Decoding – Overview

- Compared to missing data decoding, SFD
  - Takes a set of fragments instead of missing data masks
  - Outputs optimal “word string” + “segregation hypothesis”
Definition of Fragments

- **What are fragments?**
  - Spectro-temporal regions where the dominated energy is likely to have originated from a single source (not necessarily speech)
  - Sound components are grouped into fragments according to local correlations (primitive grouping)
  - This bottom-up processing is data-driven

- **Identities of fragments are not necessarily known**
  - Decided when top-down information (e.g., speech recognition models) is available
  - More on this later
A Visual Example of SFD

- Recognising “SPEECH” from noisy data
  - Consider all possible MD masks (bottom-up grouping)
A Visual Example of SFD

- Recognising “SPEECH” from noisy data
  - Consider all possible MD masks (bottom-up grouping)
- Identify fragments according to local correlation
A Visual Example of SFD (2)

- Consider all possible Missing Data masks
- Final segregation determined with top-down knowledge
A Visual Example of SFD (2)

- Consider all possible Missing Data masks
- Final segregation determined with top-down knowledge
The Search Problem

- Search space in SFD (2D)
  - Consider all possible fragment combinations
  - Also all possible word sequences given a segregation hyp.
- # of segregation hypotheses to be considered
  - Grows exponentially with time: $2^M$ where $M =$ the total number of fragments
  - $2^{13} = 8,192$
  - $2^{20} > 1$ million
  - $2^{64} = ?$
• Barker et al. (2005) developed an efficient algorithm
  • A pair of hypotheses will become identical after the offset of the last fragment by which they differ
  • The maximum number of segregation hypotheses is now $2^N$, where $N =$ the number of fragments in parallel at each frame
  • $N$ varies with time but is typically less than 4
SFD Overview – Revisited

- Combines bottom-up & top-down processing
- Simultaneously identify speech evidence and recognise speech
- Inspired by the account of auditory scene analysis
- SFD results here employ fragments generated by simply breaking the MD mask into 4 bands
Fragment Generation

- Primitive grouping based on Auditory Scene Analysis
  - Pitch (harmonicity), common fate, onset synchrony, direction ...
- Can also use MD mask estimation techniques
...we distinguish between the two sounds primarily by their difference in pitch – Hamilton, 1912
Harmonicity Based Approach

- Ma et al. (2007) generate fragments based on harmonicity
  - Based on tracking pitch of simultaneous sources
  - Makes use of the autocorrelogram to analyse harmonicity

Ma, Green, Barker and Coy, (2007) Speech Communication, 49(12): 874-891
Autocorrelogram (ACG)

- Autocorrelogram – a visual display of sound periodicity
Autocorrelogram (ACG)

- Autocorrelogram – a visual display of sound periodicity
  - 3D function mapping frequency, autocorrelation delay (lag), and time to the amount of periodic energy
Tree-Like Structures in ACG

- All frequency channels response at the F0 and this can be emphasised by the summary ACG
- Each channel also responds to the harmonic component that is closest to its centre frequency
- This produces symmetric tree-like structures appearing at intervals of pitch period

Delay = 4 ms
F0 = 250 Hz
Tree-Like Structures in ACG

- Simultaneous sources cause gaps in tree stems

![Graphic showing comparison between clean female speech and female speech mixed with male speech](image)
Spectral Grouping

Correlogram - male/female speech mixture

Gabor cosine operator

Summary correlogram

Summary - male source only

Summary - female source only
Temporal Grouping

- Pitch tends to vary little in a short time window (20ms)
- Form pitch track segments from local pitch estimates
  - Potentially overlap and extend through time
- A pitch track is used to recruit spectral segments to form a fragment
Demonstration of Fragments

- Two-speaker mixture
- Oracle segmentation
- Pitch candidates
  - Solid lines: ground-truth
  - Circles: pitch estimates
- Formed fragments
SFD on Simultaneous Speech

- Speech materials spoken by 34 English speakers (Grid corpus)
- Two talkers artificially mixed at a range of Signal-to-Noise Ratios
  - Female/Male, SNR = -6 dB: ; SNR = 0dB: ; SNR = 6 dB: ; Clean: 
  - Two male speakers, SNR = 0dB: ; Clean: 
- Performance close to human listeners

Comparing MD and SFD

- Grid corpus mixed with domestic noise (CHiME corpus)
  - Accurately replicates natural contamination (various noises plus reverberation)
  - Noise selected at the desired and naturally occurring levels
- Variable noise in low SNRs vs. stationary noise in high SNRs
  - Dry: ; Reverberated:
  - SNR = -6 dB: ; SNR = 0 dB: ; SNR = 6 dB: ; SNR = 12 dB:
Comparing MD and SFD

- Soft missing data decoding
  - Masks derived from local SNR estimates
  - Employs an adaptive noise floor tracker
- Speech fragment decoding
  - Fragments identified by tracking pitch of simultaneous sound sources

Ma, Barker, Christensen, Green. submitted to INTERSPEECH 2010
Integrating MD and SFD

- The acoustic scene can be approximately described as
  - Slowly varying noise floor (e.g. fan humming)
  - Highly unpredictable acoustic ‘events’ (e.g. a car passing by)
- **Missing Data Decoding or Speech Fragment Decoding**
  - MD with SNR masks deals well with stationary background
  - SFD copes well with unpredictable noise
Questions?