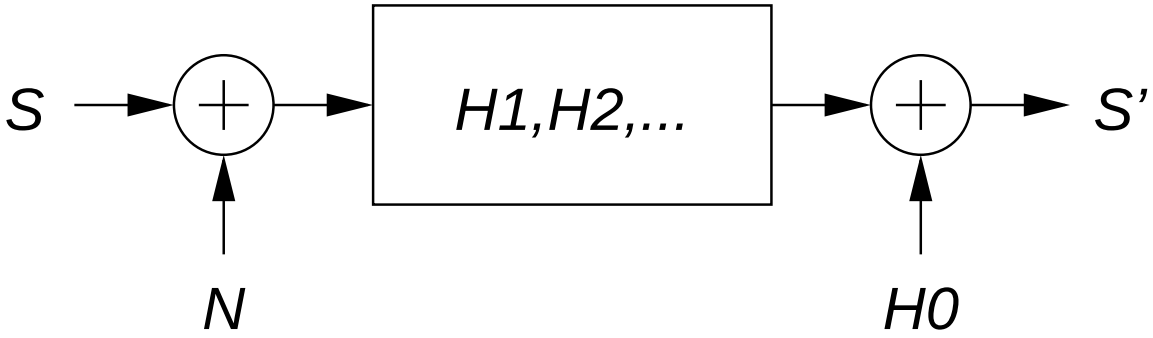
Book Chapter 8

Johan de Veth, Bert Cranen, Louis Boves: Acoustic Features and Distance Measure to Reduce Vulnerability of ASR Performance Due to the Presence of a Communication Channel and/or Background Noise

Noise model

- They consider additive background noise and channel transfer function:



$$S'(t,\omega) = H_0(t,\omega) + H_1(t,\omega)\Big[S(t,\omega) + N(t,\omega)\Big]$$

- Reliable simple model is better than unreliable complex model.

Robustness

- Not to restore the original clean speech but the **spectrum** of the clean speech.
- Actually clean speech does not exist. However, all effects similar for all utterances can be considered as part of the clean speech.
- Robustness comes into play when the effects on the speech are **variable**.
- The chapter concentrates on two things:
  1. Robustness against the channel
  2. Robustness against the background noise

Robustness against the channel – 1

- Assumptions:
  - A time-invariant channel: $H_1(t, \omega) = H_1(\omega)$.
  - Noise and the spontaneous activity of the channel can be neglected.
- $S'(t, \omega) = H_1(\omega) S(t, \omega)$
- In cepstral domain, this leads to: $c'(t, \tau) = c_h(\tau) + c(t, \tau)$.
- If training and testing is done using two different channels, we can use channel normalization methods.

Robustness against the channel – 2

- It is good to separate the following situations:
  - Conditions different for training and testing, but **constant**.
  - Conditions differ between all recording sessions.
- The paper compares three normalization methods with context-dependent and context-independent HMMs:
  - Cepstrum mean subtraction
  - RASTA filtering
  - Phase-corrected RASTA

Robustness against the channel – 3

- Cepstrum mean subtraction worked best.
- The phase-responses of the filters should be linear.
- All phase distortions interfere time-invariance and independence assumptions.
- In general, robustness methods must be compatible with the speech models.

Robustness against background noise

- The model:

$$\begin{aligned} S'(t,\omega) &= H_0(t,\omega) + H_1(\omega)\Big[S(t,\omega) + N(t,\omega)\Big] \\ &= H_1(\omega)S(t,\omega) + U(t,\omega) \end{aligned}$$

- Three approaches:
  1. Clean the features.
  2. Adapt the models.
  3. Adapt the distance computation in the search.

Cleaning the features

- If the noise is quasi time-invariant, it is possible to use spectral subtraction:
  - Estimate the noise spectrum $U(\omega)$.
  - Subtract it from the noisy input $S'(t, \omega)$.
- Needs a reliable way to estimate background noise.
- Can be combined with other methods.

Adapting the models

- Noise known:
  - Simplest to train the models in that environment.
  - Or the noise can be added artificially.
- Noise not known, but is time-invariant:
  - We can use Parallel Model Combination (PMC)
  - Separate models for different types of noise.
  - During the recognition the best combination of speech and noise is computed.
- Noise is also time-variant:
  - Much less observations for estimating $U(t, \omega)$.
  - Thus $U(t, \omega)$ has to be simpler.

Adapting the distance computation

- Noise adds uncertainty to some feature values.
- Traditional idea:
  - Detect time-frequency regions which are dominated by $U(t, \omega)$.
  - Recognition should be based on the feature values least affected by the noise.
  - How to detect noisy values?
- Another approach: Alter the distance function so that unlikely feature values have a smaller effect.

Robust local distance function

- Usually the emission probability is:

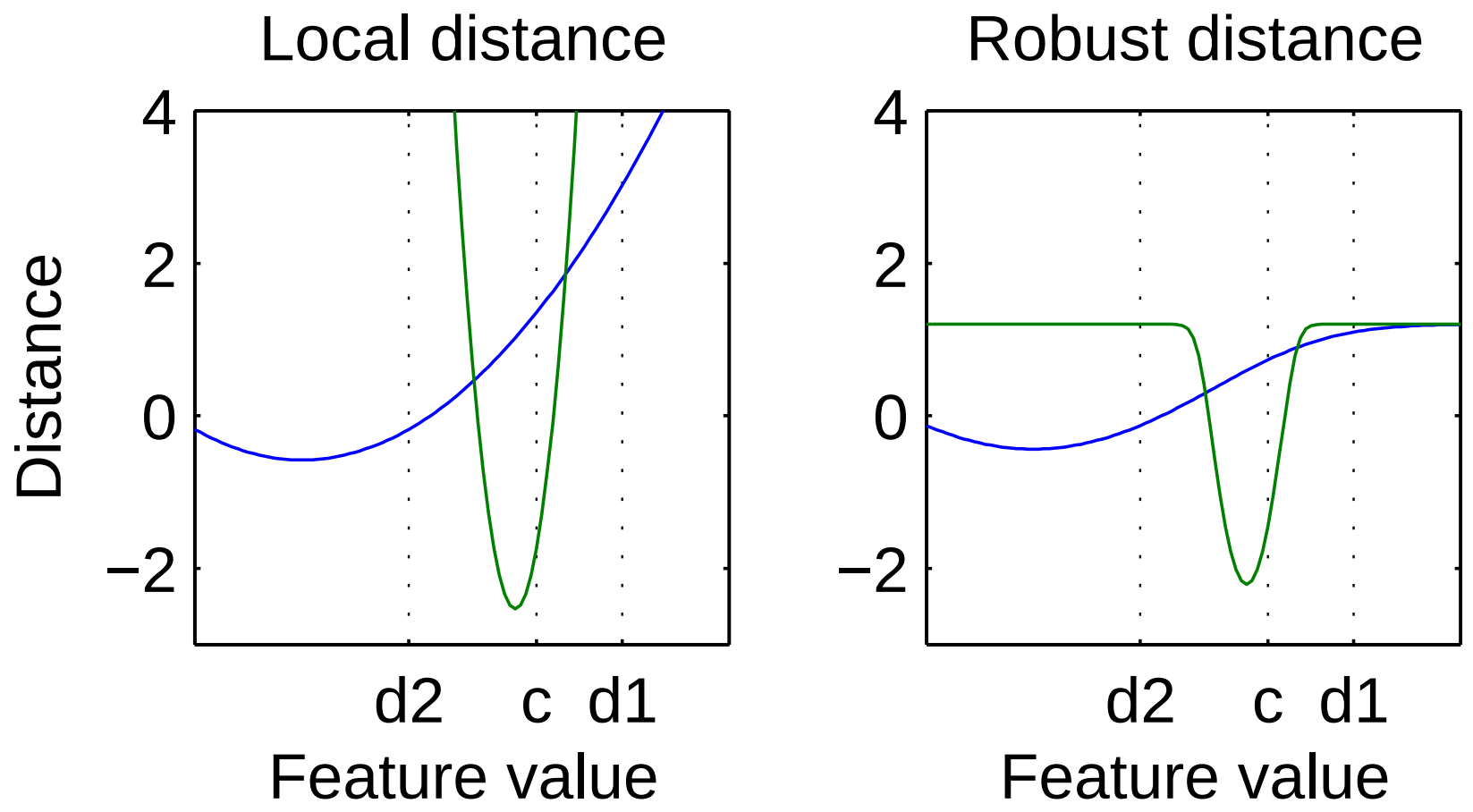
$$d_{\text{loc}}(S_i, x(t)) = -\log\left[\sum_{m=1}^{M} w_{im} \prod_{k=1}^{K} G_{imk}(x_k(t))\right]$$

- Robust local distance is:

$$d_{\text{robust}}(S_i, x(t)) = -\log\left[\sum_{m=1}^{M} w_{im} \prod_{k=1}^{K} \left[(1-\epsilon)G_{imk}(x_k(t)) + \epsilon p_0(x_k(t))\right]\right]$$

- $p_0(x)$ can be uniform distribution.

Robust local distance function



Experiments

- Connected digits recognition.
- Robust distance was compared to conventional distance.
- Results improved 0–20 %-units depending on the signal-to-noise ratios.
- Different feature transformations were also tested.
- Results were better if noise was not smeared.

Conclusions

- Channel and background noise based on a mathematical model.
- Reliable estimates are important even if it requires a simple model.
- Speech recognition modules must not violate each other's assumption.
- Noise often affects only some of the feature values $\rightarrow$ robust local distance function.
- Feature value transformations may smear the noise.