Tutorial Outline

- Probabilistic formulation LISTA
- Probabilistic mappings
- Improved modeling for synthesis

$f(s|m,y,D)$

Kleijn Henter Petkov
KTH – Royal Institute of Technology
Stockholm

$f(s|m,y,D)$

- Speech transmission $f(s|m)$
- Speech synthesis $f(s|m,D)$
- LISTA $f(s|m,y,D)$
  - s is the speech output signal
  - m is the text message or the speech input
  - y is the context signal(s)
  - D is the data base containing knowledge

s

- Is the speech output
- Practical (*is this right?*):

$$f(s|m,y,D) = \int f(s|c)\, f(c|m,y,D)\, dc$$
$$= \int f(s|c_1,c_2)\, f(c_1|m)\, f(c_2|y,D)\, dc_1\, dc_2$$

  - c_1 is "context-neutral" control
  - c_2 is controlling context indicator = *deviation from neutral*
- c is control
  - Formants, pitch, speech activity, speaking rate, power, spectral tilt, emphasis, repeat
- $f(s|c)$ deterministic (sinusoidal model) or probabilistic (autoregressive parametric model)

message synthesis already "internalized"

LISTA database

m

- m is the message for synthesis
  - The “text” of TTS
- m is the input speech for speech modification

y

LISTA "synthesis" database

- y is the context input signal(s)
  - Audio, video, separated from other signals or not
- Context *features* g:

$$f(c_2|y,D)=\int f(c_2|g,D)\, f(g|y,D)\, dg$$

LISTA "analysis" database + existing knowledge

- "Nice" context features:

$$f(c_2|y,D)=\int f(c_2|g,D_s)\, f(g|y,D_a)\, dg$$

- Typical physical-environment features
  - Car, office, restaurant, meeting environments, dark clouds(?)
- "Mirrored" features: similar to controlling context indicators:
  - Speech activity, speaking rate, power, average spectral tilt, emphasis, repeat, {"uhh", "hmm", laughing, coughing}, stress, facial expression, more?
- Signal to feature mapping:
  - Physical environment → *get/check KTH classifier*
  - "Uh", "hmm", laughing, coughing → event detection
  - Speaking rate → trivial if recognition system used; measure separately?
  - Facial expression → non-trivial and not in proposal; expertise *was* at KTH

D

- D_a is the *LISTA analysis context data base*

- Analysis context features g:

  $f(c_2|y,D) = \int f(c_2|g)\, f(g|y,D)\, dg$

  – We must model $f(g|y.D)$; approximate it as: $f(g|y,\mu,\theta)$

  – Must select model family, μ, perhaps parameters θ

- Needed:
  1. Labeled database
  2. Probability distribution model family/families

Selecting a Model Family: Frequentist Approach

- For each family select the ML parameter set
- Select the family with the highest ML score

- Approximates Kulback-Leibler divergence:

$$KL(f, \hat{f}) = \int f(x)\log(f(x)/\hat{f}(x|\theta))\, dg$$
$$= \int f(x)\log(f(x))\, dg - \int f(x)\log(\hat{f}(x|\theta))\, dx$$

likelihood

  - Sum log likelihoods = multiply likelihoods ≅ multiply likelihoods of independent observations in database

- Variations: Akaike information criterion, Takeuchi information criterion, pseudo-likelihood ratio tests, etc.

Selecting a Model Family: Bayesian Approach

- Find family $f(g|y,\mu,\theta)$ to match $f(g|y,D)$
- In the Bayesian approach model parameters have a distribution, e.g., $f(\theta|x)$, $f(\theta)$, $P(\mu|x)$.
- Evidence: $P(\mu|D) = P(D|\mu)\ P(\mu) = \int f(D|\theta)\ f(\theta)\ d\theta\ P(\mu)$
- Select family with highest posterior probability
  - Is averaging over all models in family reasonable?
- Variational Bayes: approximate $P(\mu|D)$ by $Q(\mu)$; often can optimize $Q(\mu)$ iteratively by selecting the "form" of $Q(\mu)$.

Why doesn’t it work

- Do you map the essence or the noise?

- Caricature it
- Refine model definitions

Summary

- The world from WP2 perspective
- Must define
  - Control variables
  - Features
- Need databases
- Need mappings (conditional distributions)
  - Signals to features (WP2)
  - Features to control; default, modified (WP2)
  - Control to speech (WP3, WP4)
- For all mappings
  - Must select frequentist / Bayesian selection
  - Select mappings
- Improve it by caricature