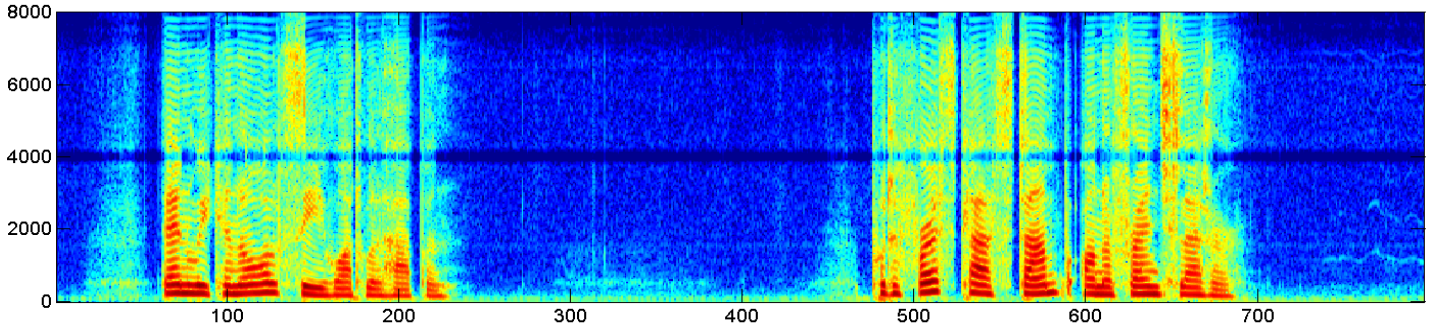
Summer School of Information Engineering
29 June – 3 July, Bressanone (BZ), Italy

Investigations about the optimization of artificial bandwidth extension

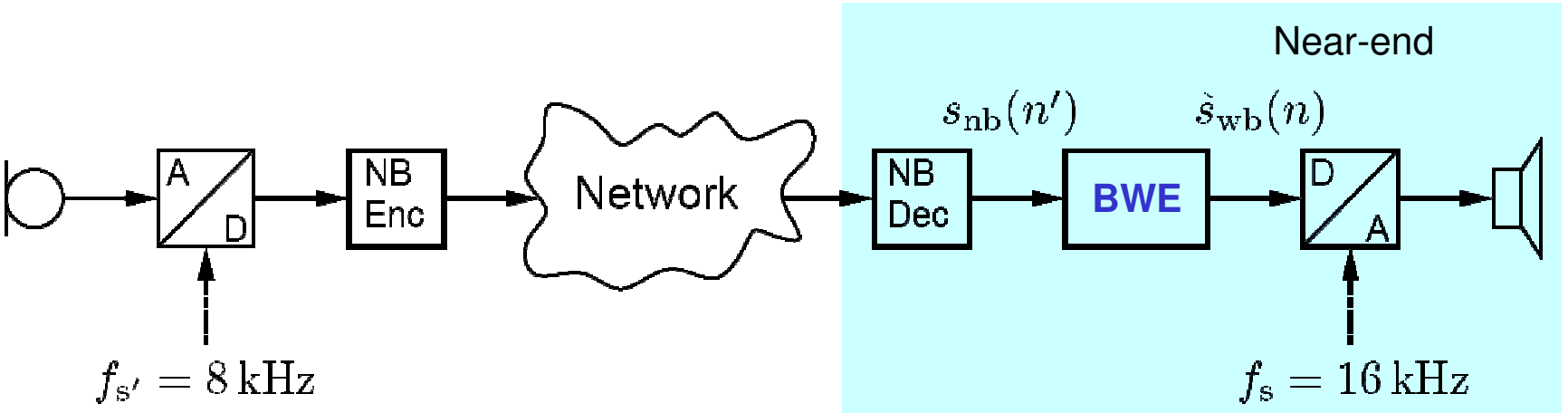


Michele Sanna
Dipartimento di Ingegneria Elettrica ed Elettronica (DIEE)
Università degli studi di Cagliari
Piazza d'Armi 09123, Cagliari, Italy
michele.sanna@diee.unica.it

Motivation (1/2)

- ✓ **Problem**:
  - Telephone signals have a limited intelligibility
    - Similar voices (father – son)
    - Syllable confusion (spelling via NATO alphabet)
    - Misunderstanding
- ✓ **Cause**:
  - The acoustic bandwidth is reduced for the 8 kHz sampling:
    - Upper limit: 3.4 kHz → reduced intelligibility
    - Lower limit: 300 Hz → low subjective quality
- **Stress and distraction** during the conversation

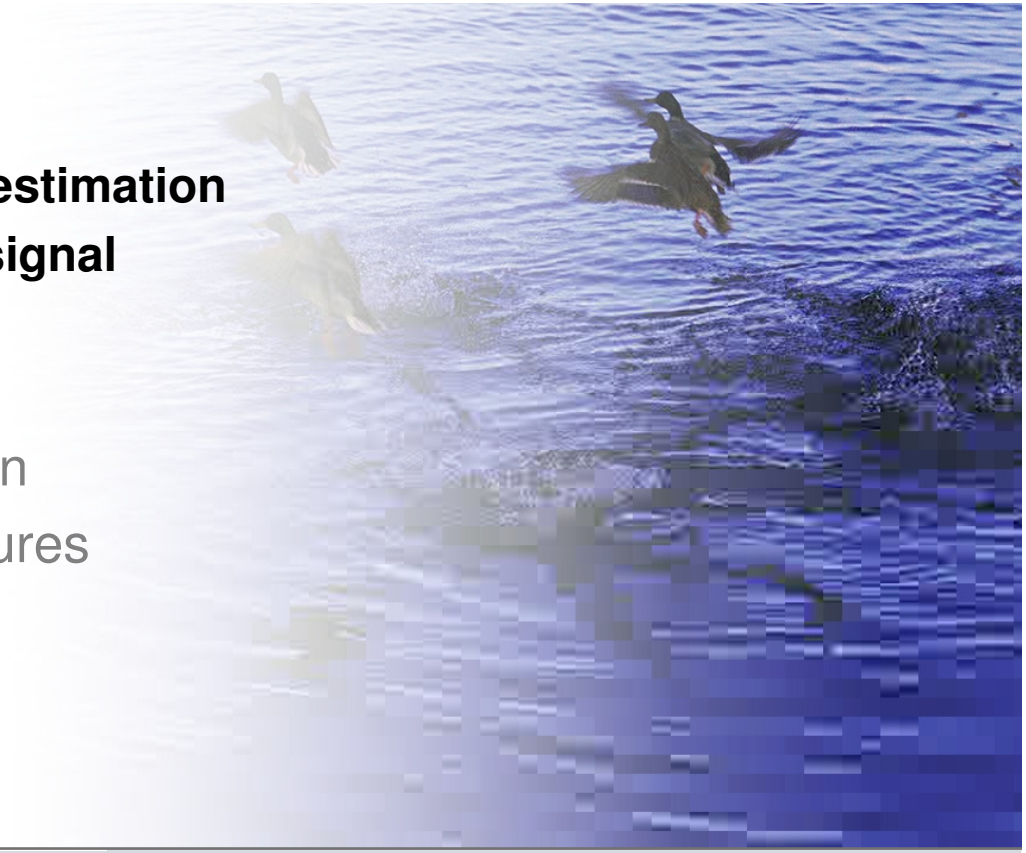
Motivation (2/2)



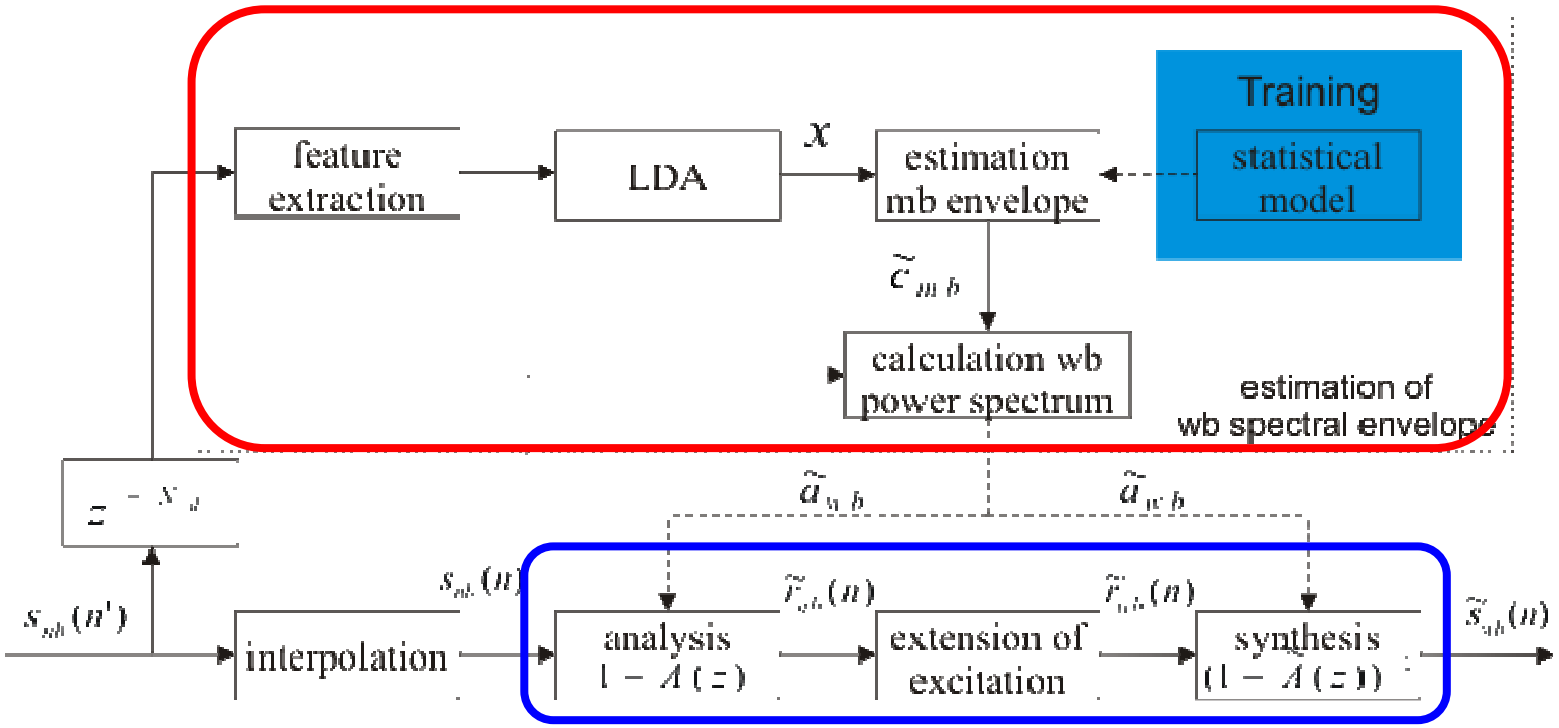
- ✓ Artificial Bandwidth Extension [ABWE]
  - Restoration of the wideband (7 kHz) without the need of additional information
  - Online processing: superior telephone quality, safer viva-voce
  - Offline processing: historical speech recordings
  - Fricative-oriented design (/s/, /z/, /f/)

Outline

1. **The ABWE algorithm**
   - **Block diagram**
   - **Statistical model: envelope estimation**
   - **Extension of the excitation signal**
   - **Result**
   - **Training**
2. Fricative-oriented optimization
3. Log-spectral distortion measures
4. Conclusions



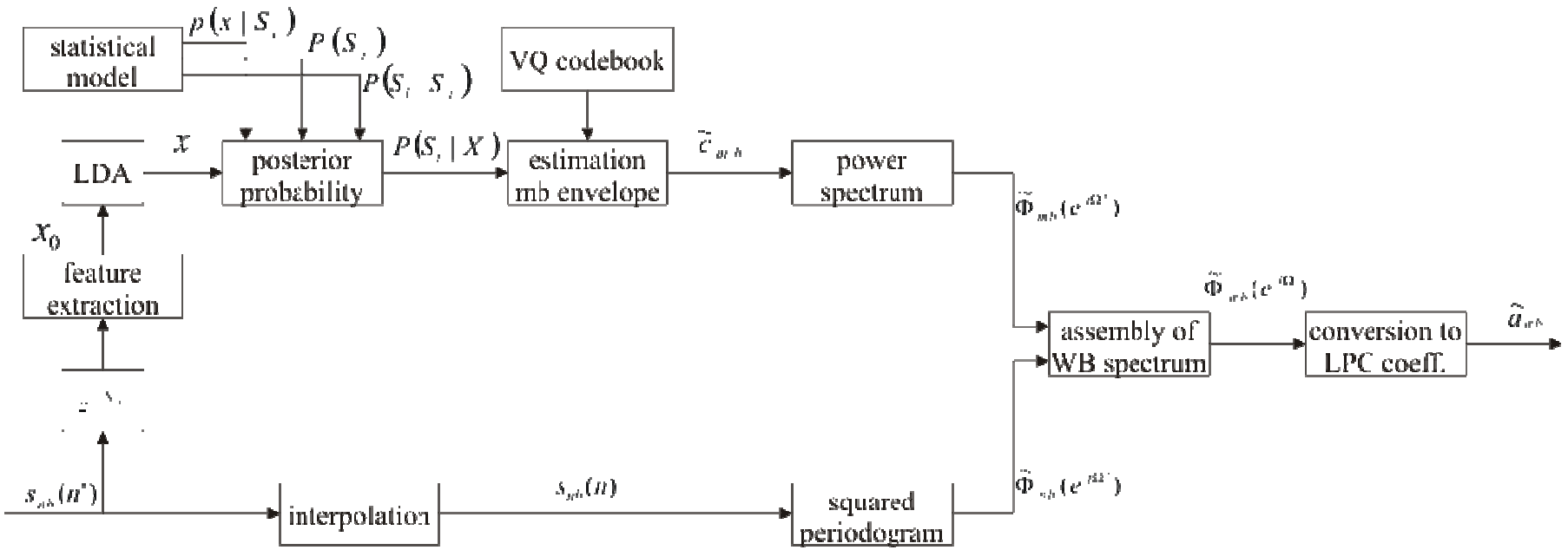
Block diagram



$s_{\mathrm{nb}}(n')$: narrowband signal
$\tilde{r}_{\mathrm{nb}}(n)$: narrowband residual
$\tilde{r}_{\mathrm{wb}}(n)$: wideband residual
$\tilde{s}_{\mathrm{wb}}(n)$: wideband speech

$z^{-N_{\mathrm{d}}}$: interpolation delay
$\mathbf{x}$: feature vector
$\tilde{\mathbf{c}}_{\mathrm{mb}}$: missing band cepstral vector
$\tilde{a}_{\mathrm{wb}}$: LPC filter coefficients

Statistical model: envelope estimation (1/3)



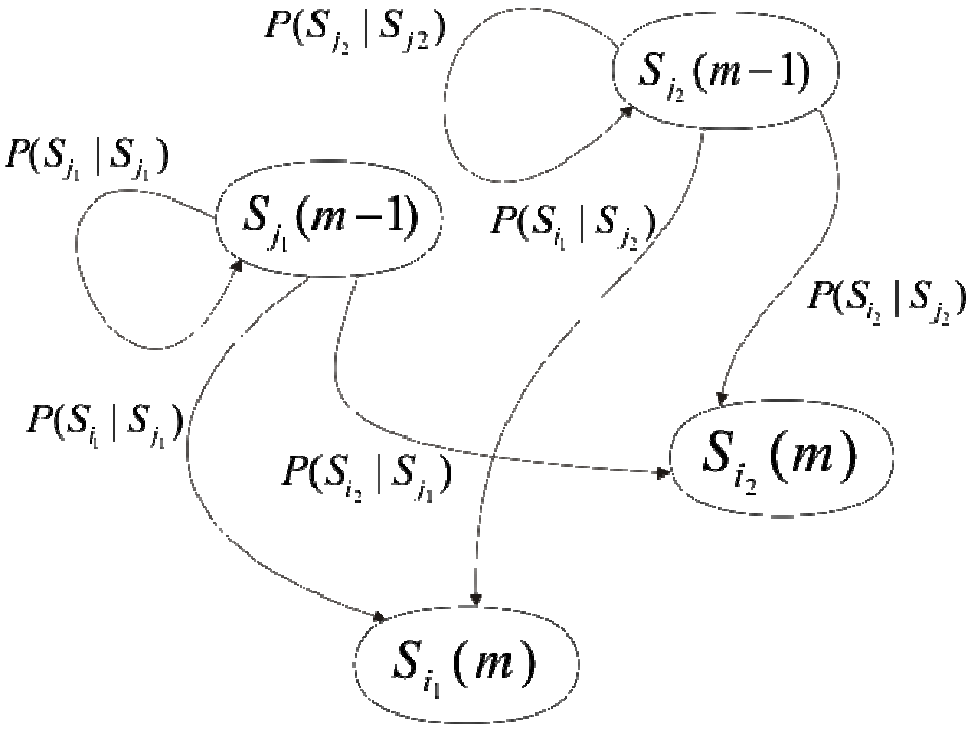
$p(\mathbf{x}|S_i)$: observation probability pdf
$P(S_i)$: state probability
$P(S_i|S_j)$: transition probability
$P(S_i|\mathbf{X})$: a posteriori probability

$\tilde{\Phi}_{\mathrm{bb}}(e^{j\Omega})$: baseband spectrum
$\tilde{\Phi}_{\mathrm{ub}}(e^{j\Omega''})$: estimated upper band spectrum
$\tilde{\Phi}_{\mathrm{wb}}(e^{j\Omega})$: assembled wideband spectrum

Statistical model: envelope estimation (2/3)

- ✓Hidden Markov Model (HMM)
  - ➢States → spectral envelopes
    $S_i \Longrightarrow \hat{\mathbf{y}}_{\mathrm{mb},i}$
  - ➢Input composite feature vector:
    $\mathbf{x}(m)$
  - ➢Transition probability:
    $P(S_i|S_j)$
  - ➢Observation pdfs
    $p(\mathbf{x}|S_i)$
  - ➢A posteriori probability:

$$P(S_i(m)|\mathbf{X}(m)) \Rightarrow \text{Bayes Eq.}$$



Statistical model: envelope estimation (3/3)

- ✓Soft-decision:
  - ➢Minimum mean square error (MMSE):

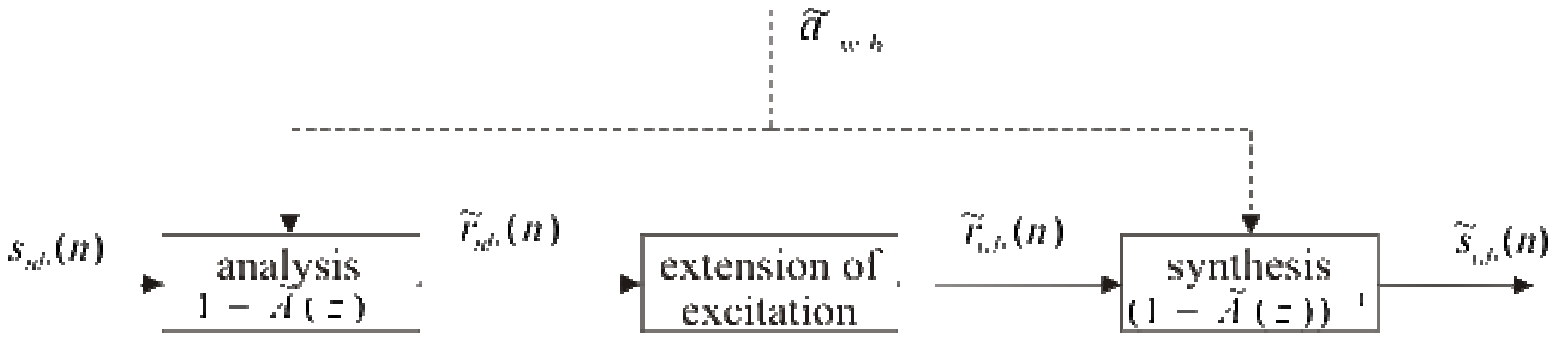
$$\tilde{\mathbf{y}}_{\mathrm{mb,MMSE}}(m) = \sum_{i=1}^{N_\mathrm{s}} \hat{\mathbf{y}}_{\mathrm{mb},i} P(S_i(m)|\mathbf{X}(m))$$

- ✓Assembly of the wideband spectrum
  - ➢Baseband periodogram (0 – 4 kHz)
  - ➢Upper band estimated (4 – 8 kHz)

- ✓LPC (Linear Predictive Coding) filter coefficients

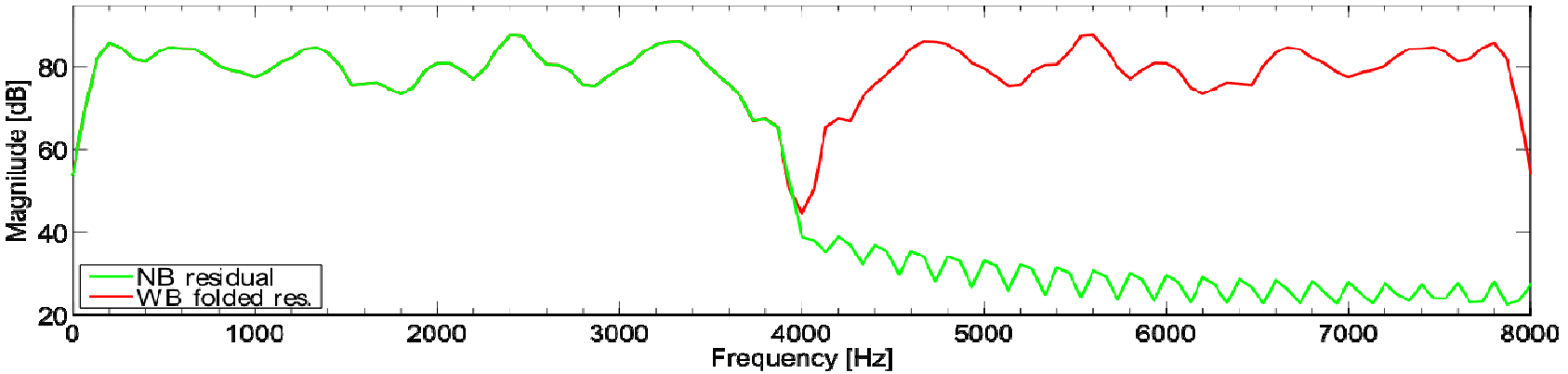
$$\tilde{\Phi}_{\mathrm{wb}}\left(e^{j\Omega}\right) \rightarrow \tilde{\mathbf{a}}_{\mathrm{wb}}$$

Extension of the excitation signal (1/2)



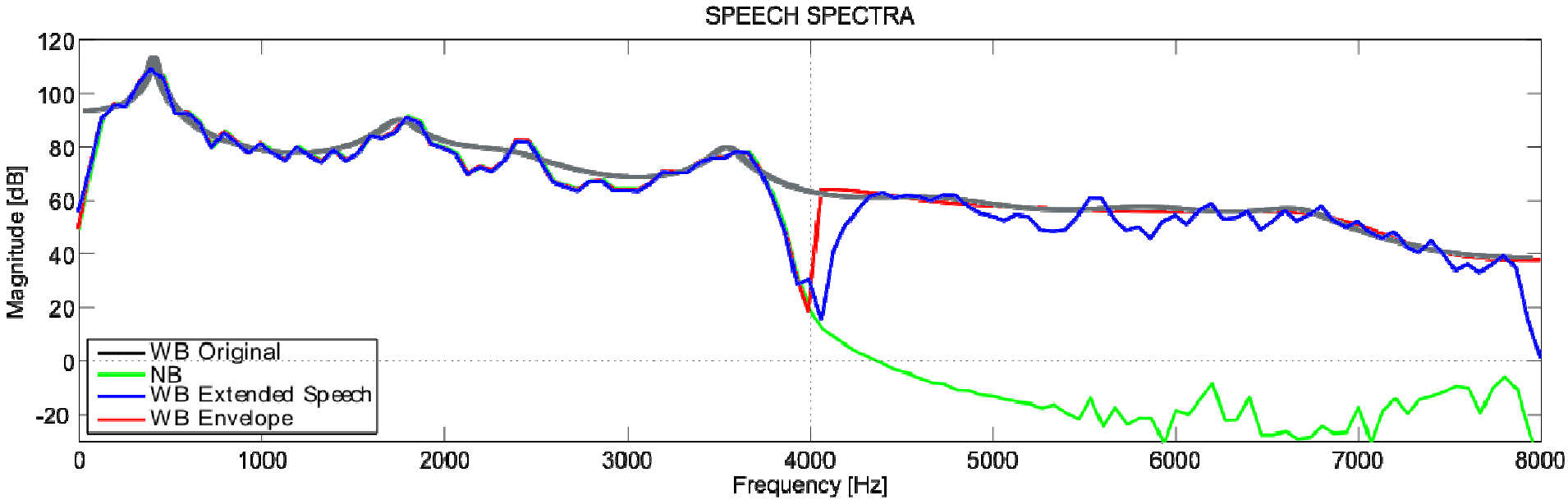
- ✓Analysis: Auto-regressive (AR) filter $1 - \tilde{A}(z) = 1 - \sum_{\nu=1}^{N_p} \tilde{a}_\nu z^{-\nu}$
  - ➢Residual signal $\tilde{r}_{wb}(n)$
- ✓Extension of the residual to 8 kHz
- ✓Synthesis: linear prediction (LP) filter $\left(1 - \tilde{A}(z)\right)^{-1}$
  - ➢Excitation signal $\tilde{r}_{wb}(n)$

Extension of the excitation signal (2/2)



- ✓Narrowband residual signal (white noise): $\tilde{r}_{\mathrm{nb}}(n)$
- ✓Folding around 4 kHz:
  - ➢0 – 4kHz ⟶ 0 – 8 kHz
- ✓Excitation of the synthesis filter: $\tilde{r}_{\mathrm{wb}}(n)$
  - ➢Wideband voice!

Result



4. Synthesized wideband signal: **Baseband transparency!!**

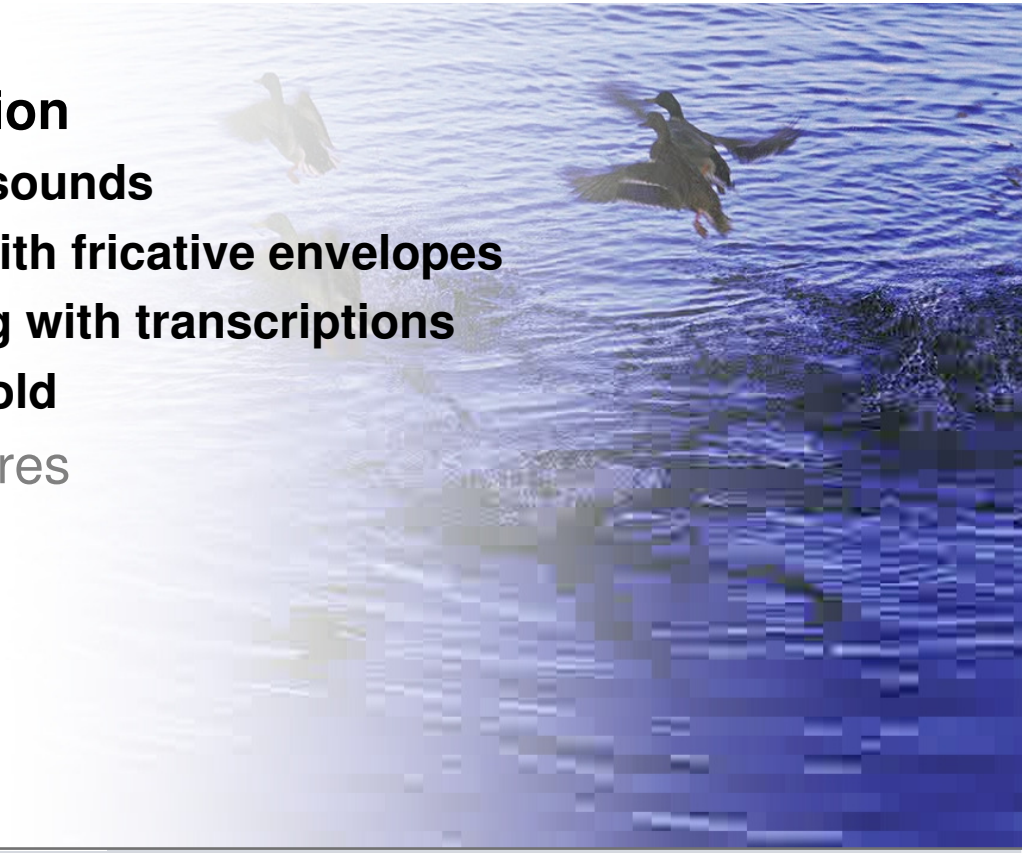
Training phase

- ✓Codebook Training
  - ➢Selective Linear Prediction (SLP) analysis, upper band
  - ➢LBG vector quantization (VQ)
    - ❖**Codebook (CB – 16 states S_i)**

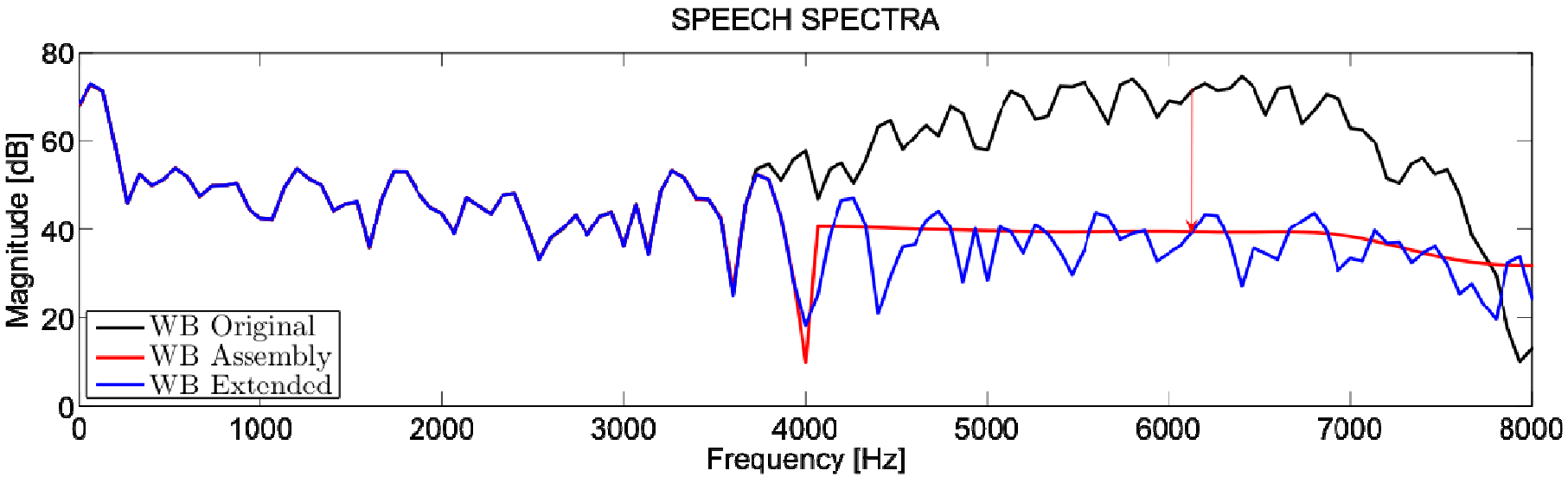
- ✓Training of the statistical model (HMM)
  - ➢State statistics:
    - ❖**State probabilities** $P(S_i)$
    - ❖**Transition probabilities** $P(S_i|S_j)$
  - ➢Likelihood statistics
    - ❖**Observation pdfs** $p(\mathbf{x}|S_i)$

Outline

1. The ABWE Algorithm
2. **Fricative oriented optimization**
   - **Underestimation of fricative sounds**
   - **Extension of the codebook with fricative envelopes**
   - **Transition matrix and training with transcriptions**
   - **Transition probability threshold**
3. Log-spectral distortion measures
4. Conclusions



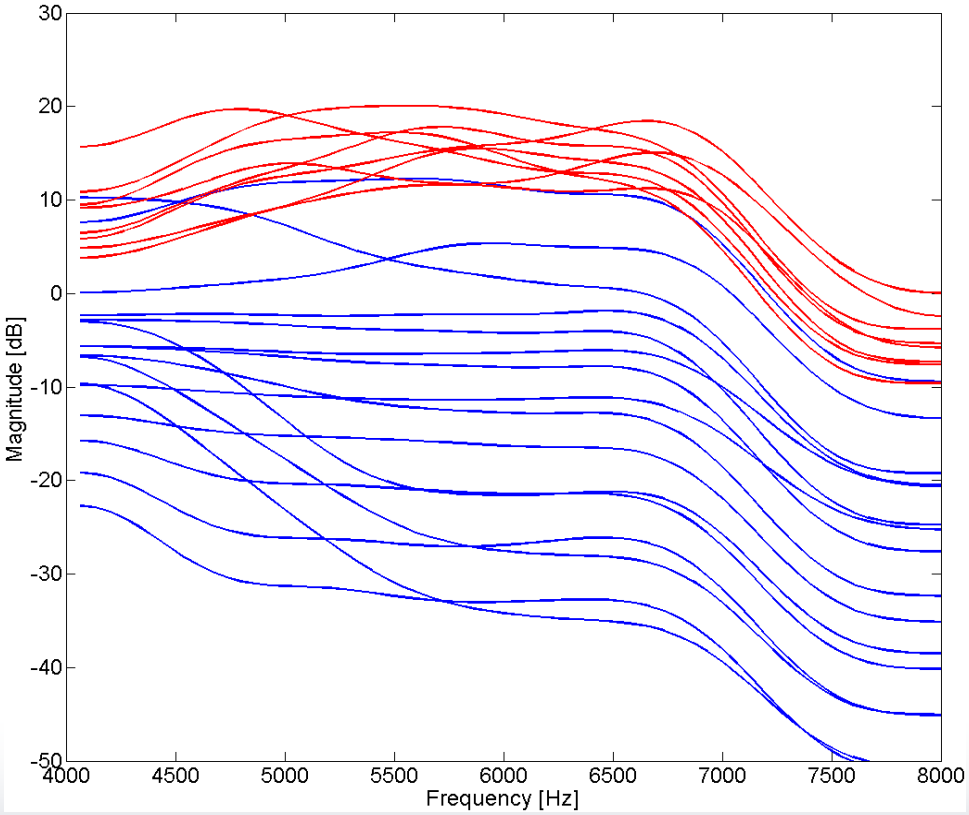
Underestimation of fricative sounds



✓High energy content in the upper band ⟶ difficult recognition

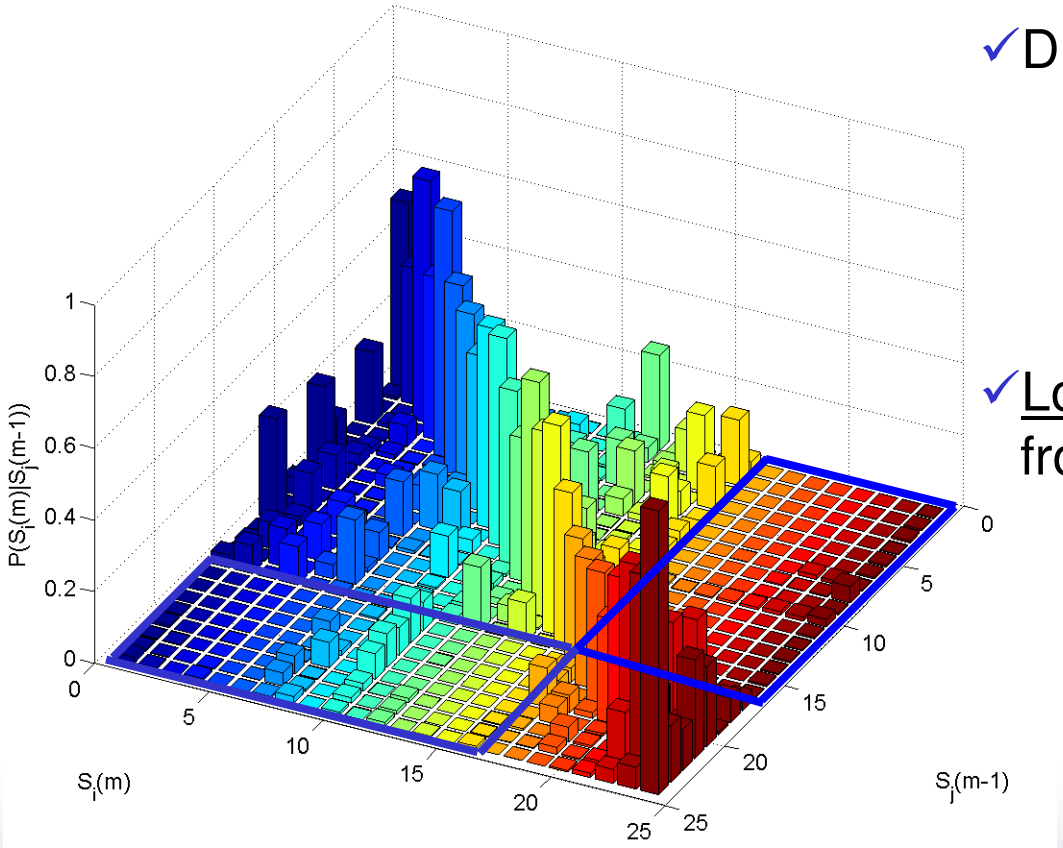
✓Inadequate envelope ⟶ energy underestimation ⟶ unfaithful sound

Extension of the codebook with fricative envelopes



- ✓ Increment of the statistical representation of the fricatives
  1. Material composed by /s/ and /z/
  2. Extraction of 8 representatives
  3. Assembly of the new states to the natural 16
  4. Extended codebook, 24 states
     - ❖ Training with transcriptions!

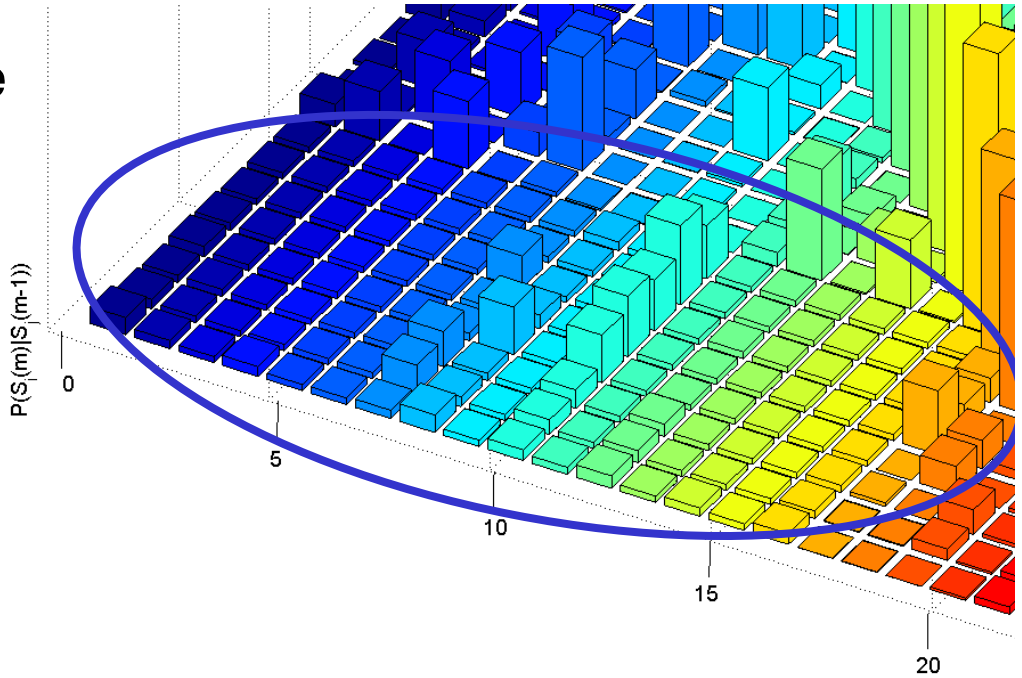
Transition matrix and training with transcriptions



- ✓During the training:
  - ➢Assignment of /s/ and /z/ to the new states
  - ➢Augmented statistics
- ✓Low transition probabilities to and from the new states
  - ➢Transitions delayed:
    - ❖Rise of the envelop (/s/ on-set)
    - ❖Decay of the envelope (/s/ off-set)
    - ❖Lisping effect and high frequency whistling

Transition probability threshold

- ✓Insufficient training material for the new states
- ✓Modification of the HMM transition probability matrix:
  - ➢Adding of 0.01 offset to both rectangles
  - ➢Renormalization
- ✓Enhancement of the transitions from and to the fricative group



Outline

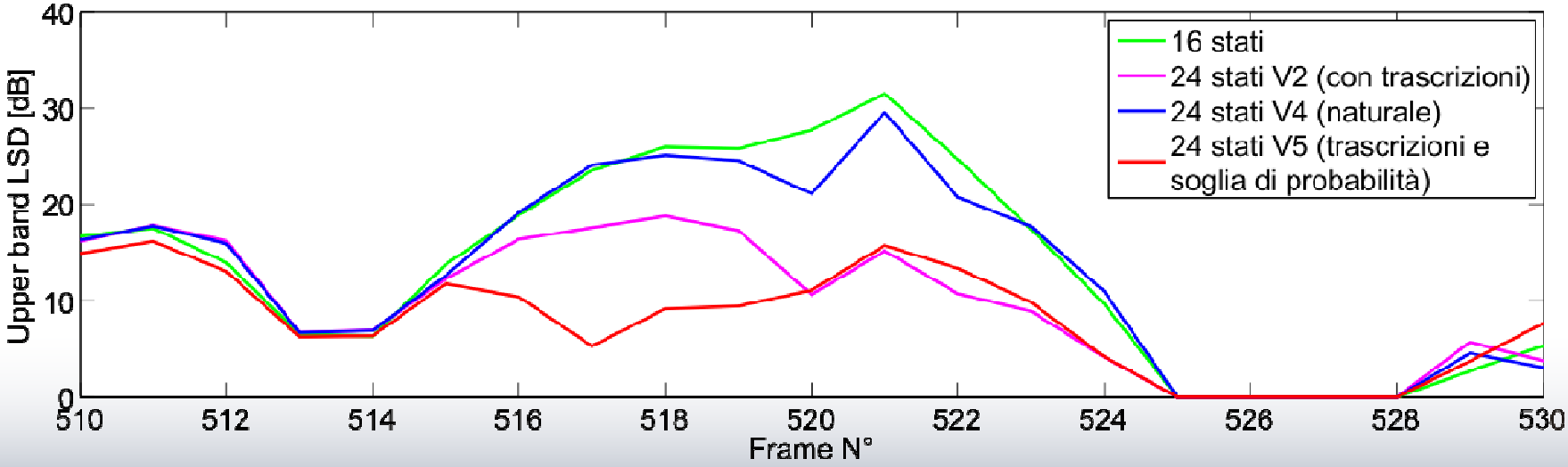
1. The ABWE algorithm
2. Fricative oriented optimization
3. **Log-spectral distortion measures**
   - **Log-spectral distance (LSD)**
   - **Mean distortion**
   - **Audio demo**
4. Conclusions

Log-spectral distance (LSD)

✓Logarithmic distance between the upper band cepstra:

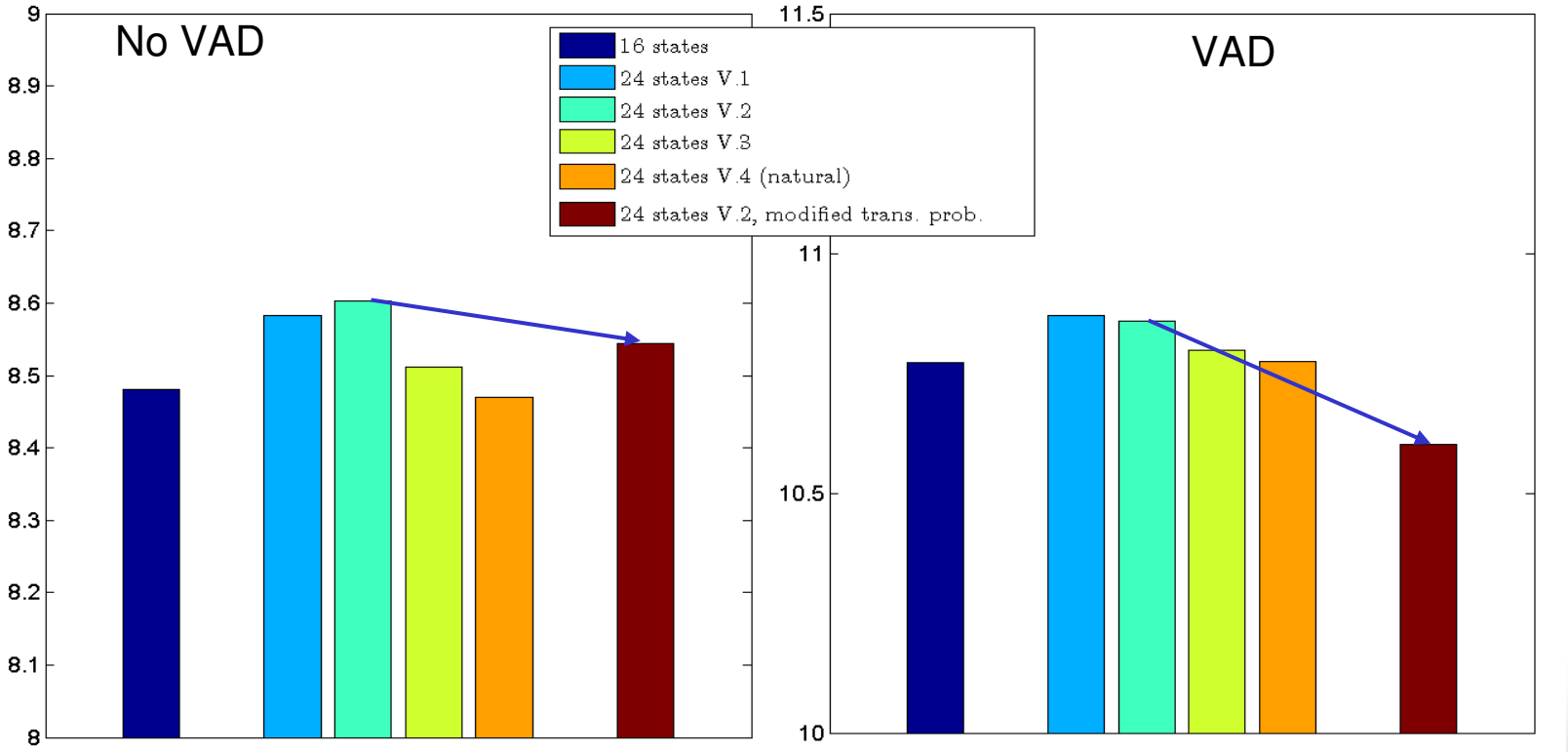
$$d_{\mathrm{lsd}} = \frac{10}{\ln 10} \cdot \sqrt{(c_0 - \tilde{c}_0)^2 + 2\sum_{d=1}^{\infty}(c_d - \tilde{c}_d)}$$

➢Example of distortion during an /s/ sound



Mean distortion

LSD measure on a database without transcriptions



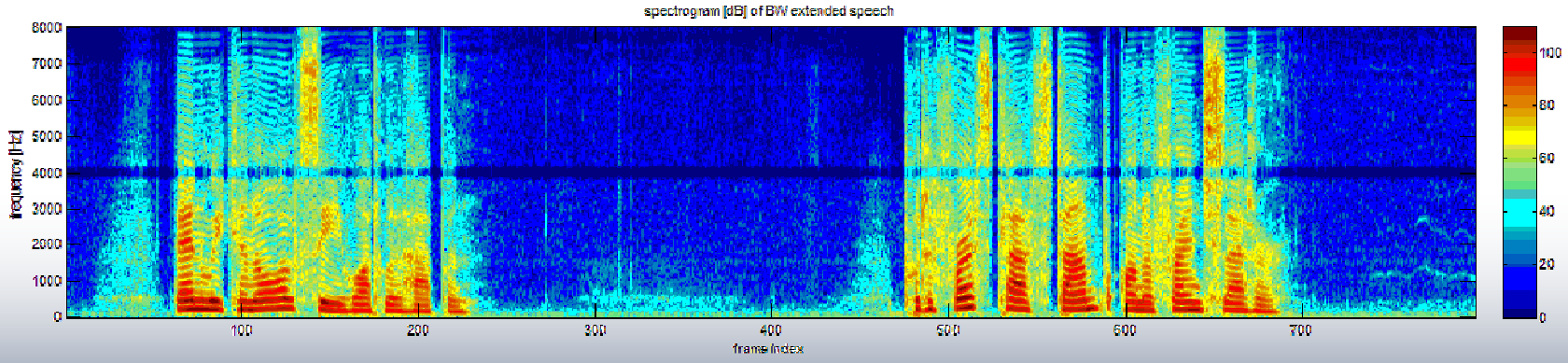
Audio demo

- "Then we can all see what will happen. Put a first class stamp on a French letter"

Narrowband speech

BWE speech - 16 states (no fricative states, natural training)

BWE speech – 24 states (8 fricative states, training with transcriptions and modified HMM transition matrix)



Summary

1. The ABWE algorithm
2. Fricative oriented optimization
3. Log-spectral distortion measures
4. **Conclusions**

Conclusions

- **Changes** to the basic implementation of the BWE:
  - Addition of fricative oriented states
  - Modifications to the training strategies
  - Modification of the HMM transition matrix

- **Result**:
  - HMM has a good reactivity
  - The reproduction of the fricatives is clearer

- **Outlook**:
  - Discriminative training using transcriptions
  - Extension under 300 kHz
  - Study of the offline processing

Thanks for the attention

Any question?

Contact:

michele.sanna@diee.unica.it

DIEE – Dipartimento di Ingegneria Elettrica ed Elettronica
Università di Cagliari, Piazza d'Armi 09123 Cagliari