

THEME 2

Speech and audio context recognition

March 27th, 2006



IC
CRIM 20 ANS

Votre **accélérateur**
technologique

Research Theme 2

Speech and Audio Context Recognition

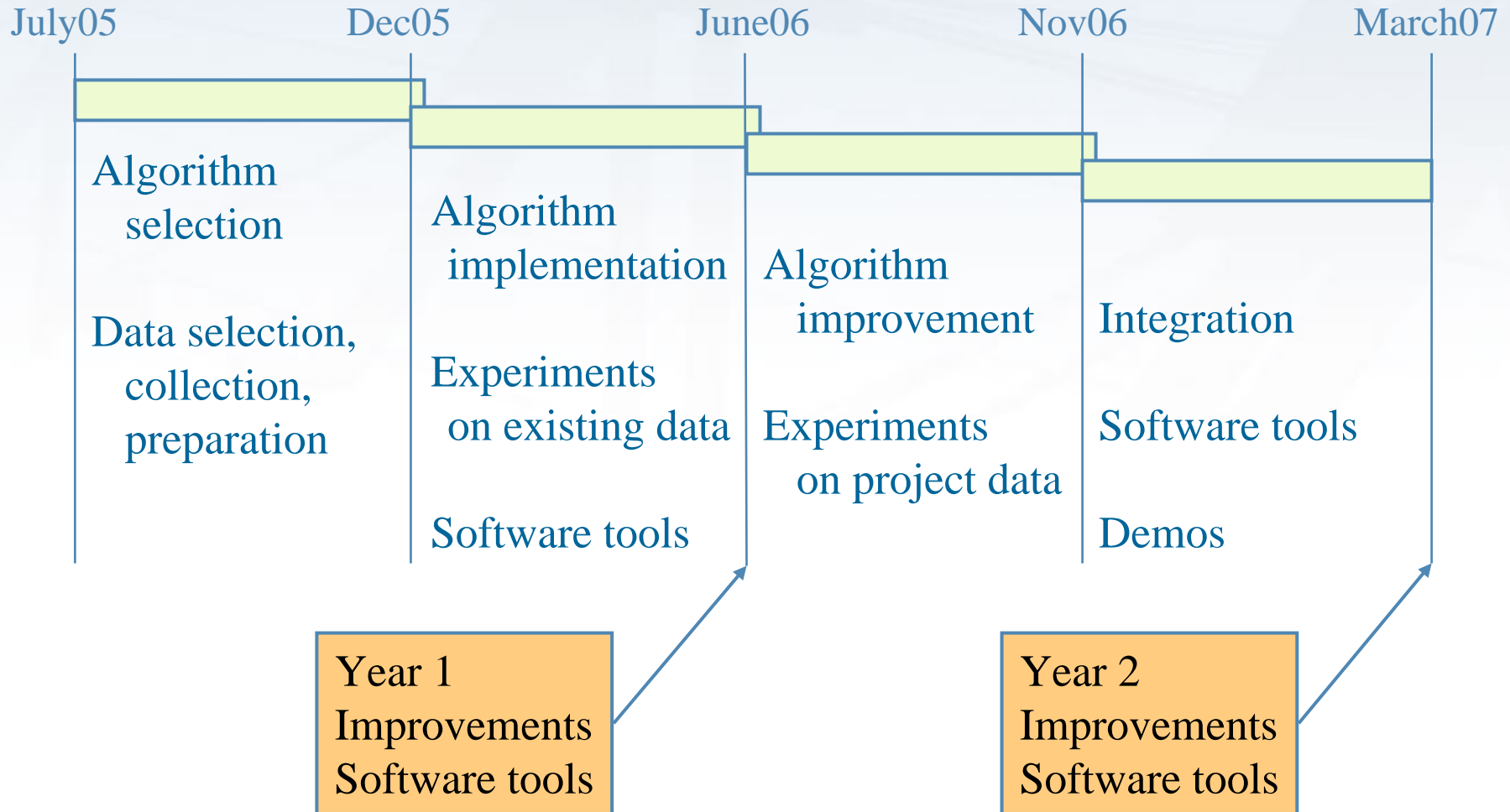
- **Theme Leader: Gilles Boulianne**
- **Research Contributors: CRIM, McGill University, École de Technologie Supérieure**
 - Research & develop speech recognition software tools addressing the most labour-intensive aspects of media production and post-production, including post-synchronization, closed-captioning, script correction, and subtitling.
- ***Project 2.1: Core Speech Technology Improvements***
- ***Project 2.2: Speech & Audio Context Recognition***

MAIN PROBLEMS

- Core speech technology improvements
 - to reach useful levels of accuracy for cultural content, speech-to-text have to be improved and adapted
- Cultural speech
 - wider ranges of topics, environments, style, language
- Audio context
 - conveys relational and emotional content not present in transcribed speech
 - needs to be identified, classified, exploited
 - to improve accuracy, to provide end-users with more complete access to cultural content

Speech and audio context recognition

PROJECT TIMELINE



Core speech technology improvements

To reach useful levels of accuracy for cultural content, speech-to-text have to be improved and adapted

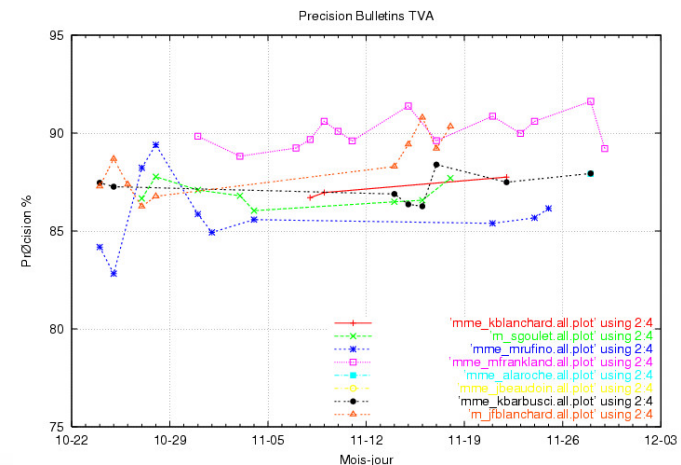
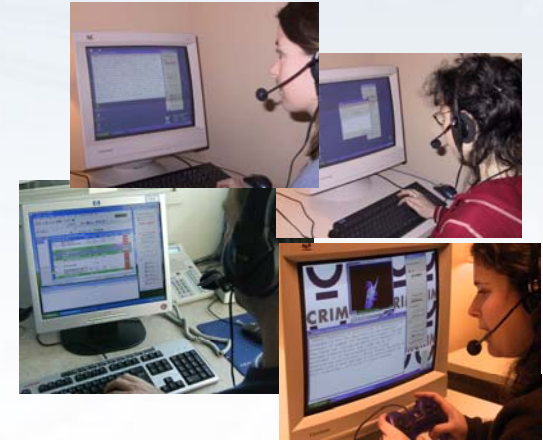
- Live TV broadcast closed-captioning multi-speaker database
- Key topics :
 - environment normalization
 - search network optimization
 - language model adaptation
 - discriminative training



Core speech technology improvements

MULTI-SPEAKER DATABASE

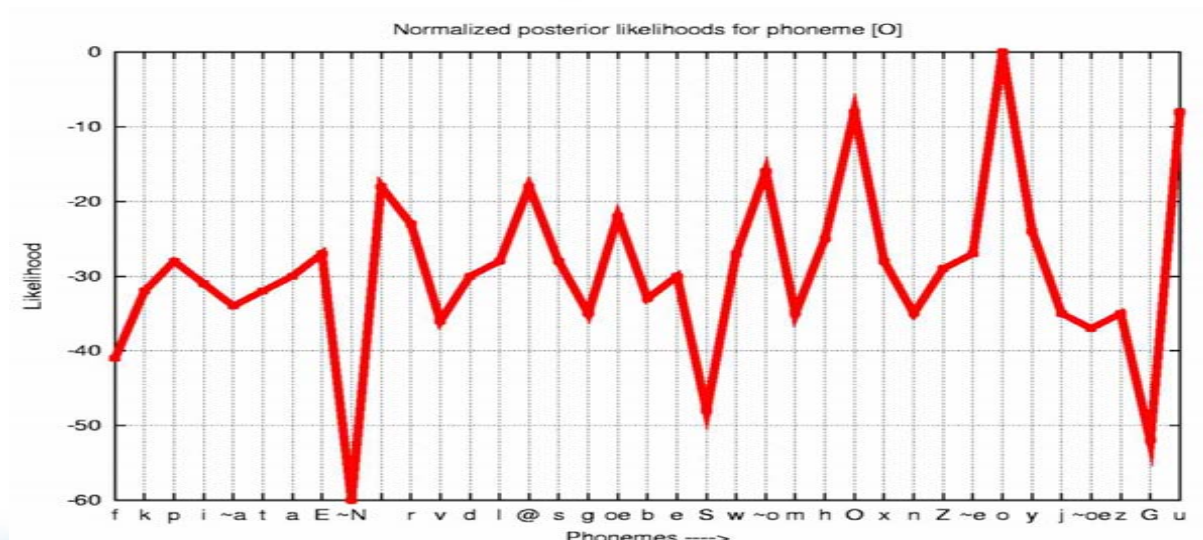
- Data collected during actual live closed-captioning at TVA (from October 2004 to October 2005) and at CRIM (from February 2005 to October 2005).
- Multi-speaker database, over 55 hours of audio data from 22 shadow speakers.
- Reference transcriptions : hand-corrected captions, topic identified from markers inserted during captioning.
- Similar database from another 51 hours, 2 shadow speakers for speaker-dependent experiments.



Core speech technology improvements

Vishwa Gupta, CRIM

- Acoustic modeling
 - Model size, parameter sharing
 - Endpointing, real-time feature normalisation
 - Multispeaker SI accuracy : from 76.1% to 82.5%
- A posteriori phoneme likelihood features
 - Paper submitted to *InterSpeech 2006*

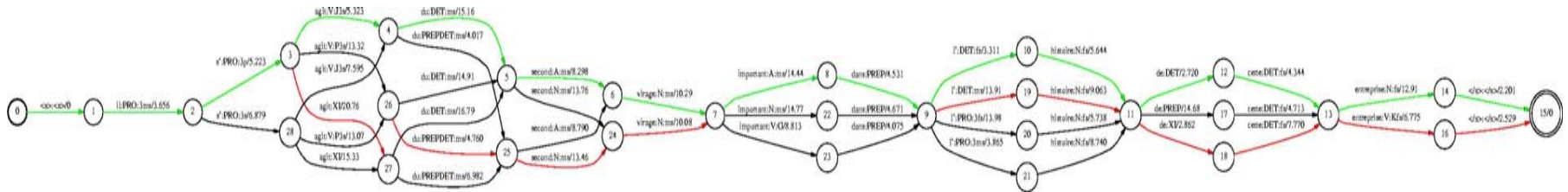


Core speech technology improvements

Maryse Boisvert, CRIM

Language modeling

- Improvements in MDE adaptation
- Vocabulary selection
- Unsupervised tagger for gender & number agreement



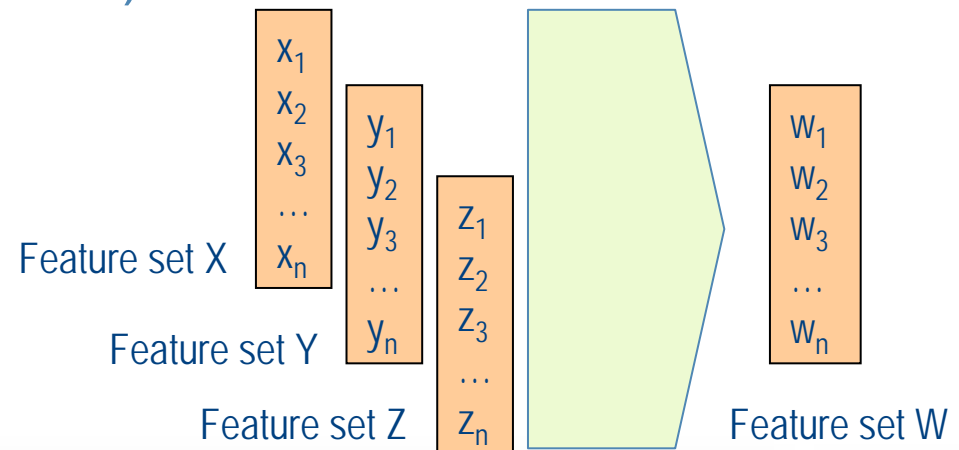
Il s'agit du second virage important dans l'histoire de cette entreprise.

Core speech technology improvements

Gilles Boulianne, CRIM



- Linear transformations of models and features
 - Linear discriminant analysis (LDA)
 - Semi-tied covariances (STC)
 - Maximum likelihood linear transformation (MLLT)
 - Extended MLLT (EMLLT)
 - Heteroscedastic LDA (HLDA)
- Discriminative criterion, dimensionality reduction
- Implementation (HLDA+MLLT)



Core speech technology improvements

Pierre Ouellet, CRIM



- Context dependence (PICs)
- Larger contexts, more complex models

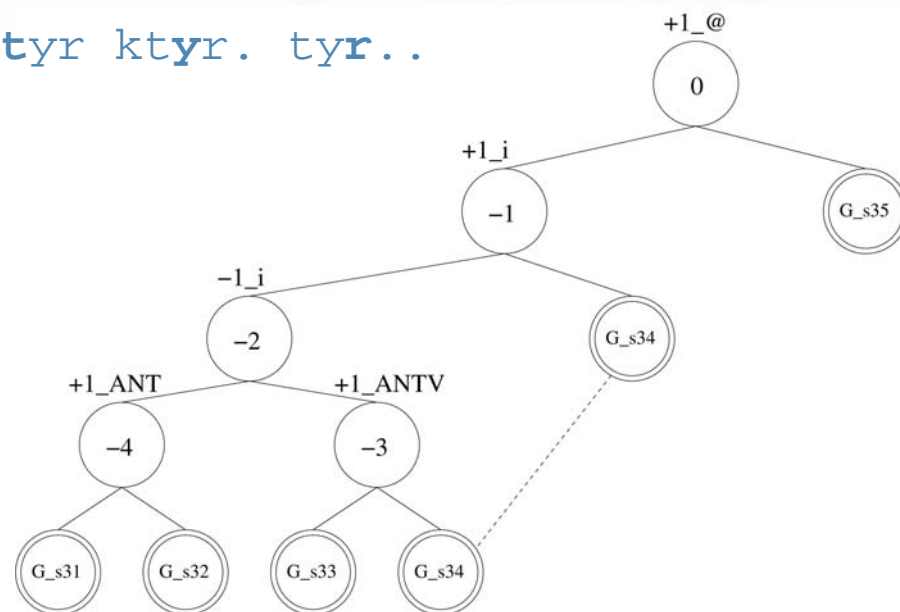
s t r y k t y r
.st str try ryk ykt kty tyr yr.
..str .stry stryk trykt rykty yktyr ktyr. tyr..

$37^1 = 37$ phonemes-in-context

$37^3 = 50$ K phonemes-in-context

$37^5 = 69$ M phonemes-in-context

$37^7 = 95$ T phonemes-in-context



Core speech technology improvements

Patrick Cardinal, CRIM

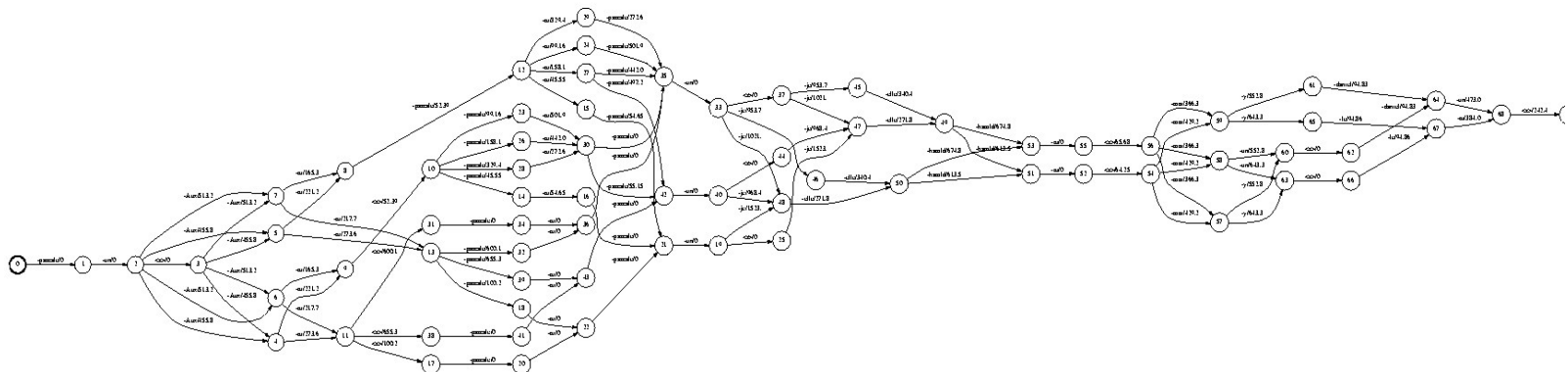


■ Discriminative training

- maximum mutual information (MMI)
- minimum phone error (MPE)

■ Implementation

- transducer-based lattice forward-backward
- model reestimation



Speech and audio context recognition

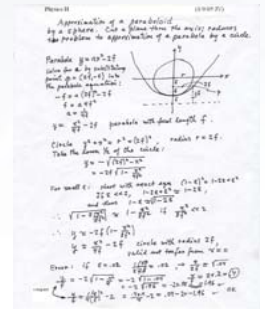
Cultural speech has wider ranges of topics, environments, style, language
Audio context conveys emotional content not present in transcribed speech

2.2.1 Cultural speech and audio context (CRIM)

2.2.2 Lecture speech (McGill)

2.2.3 Emotional content (ÉTS)

2.2.4 Sync with actor discourse (ÉTS)



Speech and audio context recognition

BROADCAST DATABASE

- whole TV shows, including music and commercials
- various broadcasters, large range of subjects
- verbatim transcriptions (3 M words)
- speaker turns and speaker identities
- 143 hours of speech from 209 speakers
- (plus 51 hours of music or commercials)



The screenshot shows a software interface for speech analysis. The top part displays a transcript with speaker names and their corresponding speech segments. The bottom part shows an audio waveform and a timeline with speaker labels.

Eve-Marie Lortie

- Colette, le mois d'août achève et ça paraît, hein. Les soirées sont fraîches.

Colette Provencher

- [b-] oui [b-], on s'habille un petit peu plus chaudement. [b] Écoutez, ce soir, c'est particulièrement frais. On n'a pas eu beaucoup de soleil sur le sud-ouest du Québec. Il y a des nuages qui ont débordé. Il y avait un système dépressionnaire qui touchait la côte est américaine et, bon, il y a des nuages qui ont débordé sur nos secteurs. Donc, e étant donné qu'il n'y avait pas de soleil.
- Les températures étaient un peu fraîches. Mais, ne vous inquiétez pas: c'est presque terminé. Le ciel se dégage au cours de la nuit prochaine. Demain, ce sera beau.
- un peu plus chaud. Tous les détails plus tard.

Eve-Marie Lortie

- À tout de suite. [b]
- Maurice Boucher subira finalement un procès séparé pour répondre aux accusations de gangstérisme et complot pour meurtre.
- La Couronne aurait voulu que Boucher soit jugé en même temps que 17 autres Hell's Angels, dans ce fameux méga-procès qu'on doit reprendre cet automne.
- Mais, la Couronne s'y est prise trop tard.
- André Jobin.

André Jobin

- La poursuite a signifié au tribunal qu'elle voulait inclure Maurice Boucher au groupe des 17 présumés [b-] membres des Hell's Angels accusés de gangstérisme, trafic de stupefiants et complot pour meurtre.

TVAO20829_22000_30

Resolution 511

report

(no speaker)	[single]	[single]-[audi 28 août en ...]	[single]-[Des guich Secur...
--------------	----------	--------------------------------	------------------------------

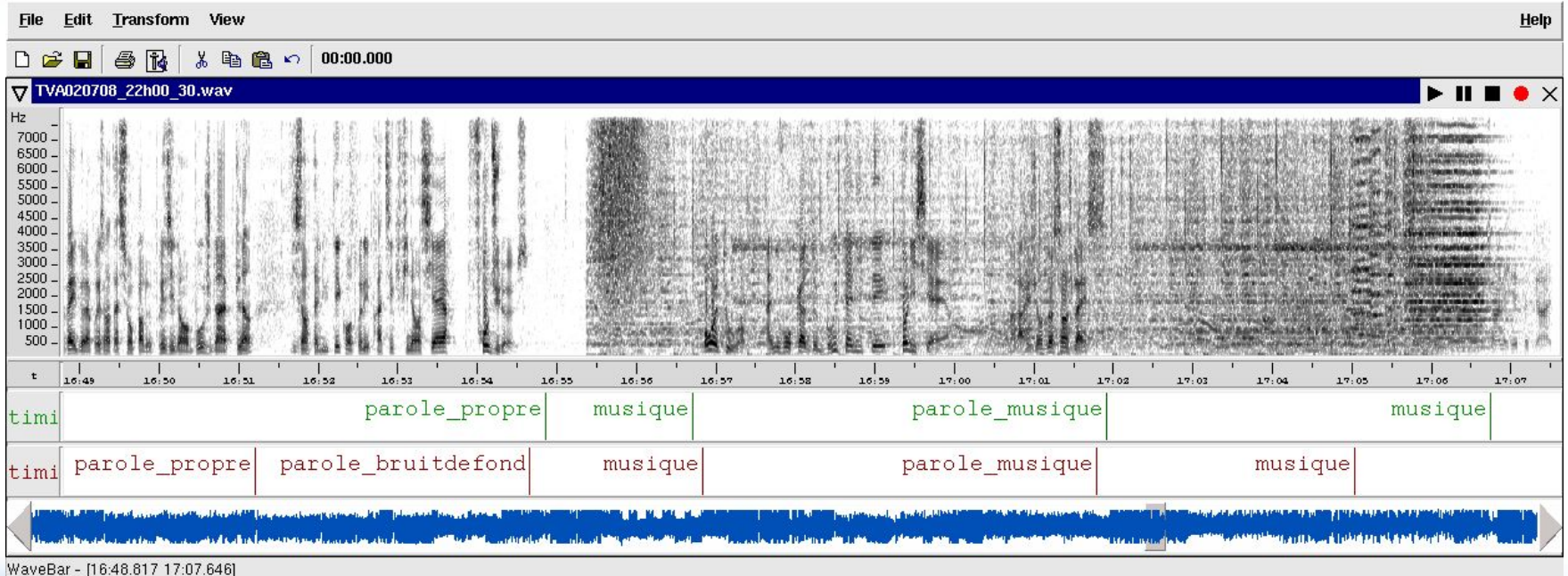
Signal shape is now available!

Speech and audio context recognition

Michel Comeau, CRIM



Acoustic segmentation and classification



WaveBar - [16:48.817 17:07.646]