Mbrola Diphone Experience

Multitel – TCTS Lab, Mons, Belgium

- 25 Languages, 50 dbas (4 german, 2 more soon)

- Mbrola database sources
  - Only %20 of the databases are recorded in TCTS Lab
  - Variety of conditions in recording

Diphone Database Building Details – Speaker Choice

- **Speaker Choice**
  - **Output quality is speaker dependent**
  - **Choice by trial error**
    - **Recording a subset just enough to synthesize a few phrases**
    - **Decode/encode synthesis of a few phrases**
    - **Detection of irregularities in speech or other problems**

- **Generalized to long corpora recording**
  - **Speakers voice should be tested with several signal processing algorithms (Mbrola team is volunteered to help)**
    - **Additionally, testing synthetic speech on telephone is also useful**
  - **Availability of speaker for long recording sessions**
  - **Speaker selection with help of signal processing tools??**

Diphone Database Building Details – Text Design

- **No universal solution**
  - **Meaningless logatoms (if letter to sound rules are obvious),**
  - **Words from lexicon (picking from a dictionary with constraints),**
  - **Phrases containing multiple diphones**

- **Similiar context helps reducing discontinuities**
  - **X p1 p2 X , X p1 p2 Y**
  - **Avoiding pitch attacks and vocal fries**

- **Generalized to long corpora recording (just the inverse)**
  - **Context variability is advantegous**
  - **Various prosodic events are advantageous (variance and controlability needs to be discussed)**
  - **Rather more complex set coverage problem**

Text Design

- **A Set Coverage Problem**
  - **Limited vs unlimited domain synthesis**
    - **What to be covered?**
- **Two approaches**
  - **Iterative corpus building**
    - **With the existance of high quality automatic segmentation and unit selection system**
    - **Corpus building tuned to speaker and the system**
  - **Speaker independent text selection**
    - **Defining the set to be covered, not everything can be covered**
      - **Phonetic coverage**
        - » **Units?? Diphones, triphones, words,…**
        - » **Context?? phoneme similiarities ?**
      - **Prosodic unit coverage**
    - **Method : Greedy is most common**

Recording (Diphone and NUU Corpora)

- **Availability**
  - **Studio shall be available for long recording period(months)**
- **Unechoic conditions?**
  - **Signals must be pure (signal processing after recording may be risky)**
  - **Maybe hard for speakers to stand conditions for long time, he/she may tend to finish as soon as possible to get out of the studio**
- **Variations in and inbetween sessions**
  - **For diphones, prompts may be used…for NUU corpora??**
- **Helping the speaker**
  - **Monitoring the process**
    - **One person for monitoring signals, one for guidance in studio**
    - **Signal processing tools for monitoring?**

Recording (Diphone and NUU Corpora)

- **Some sources of degradation**
  - **The algorithm itself introduces degradation (the amount is usually speaker dependent)**
    - **Mbrola −> phase distortion**
  - **Recorded sound characteristics**
    - **Prosodic modification degrades signal quality**
      - **Speech rate and prosodic variation are important**
    - **Amplitude variations in diphones**
      - **Equalization for diphones is rather easy**
      - **Not trivial for NUU databases**