THINKTech

Budapest University of Technology and Economics

AITIA

IMPROVED RECOGNITION OF HUNGARIAN CALL CENTER CONVERSATIONS

B. Tarján, G. Sárosi, T. Fegyó, and P. Mihajlik

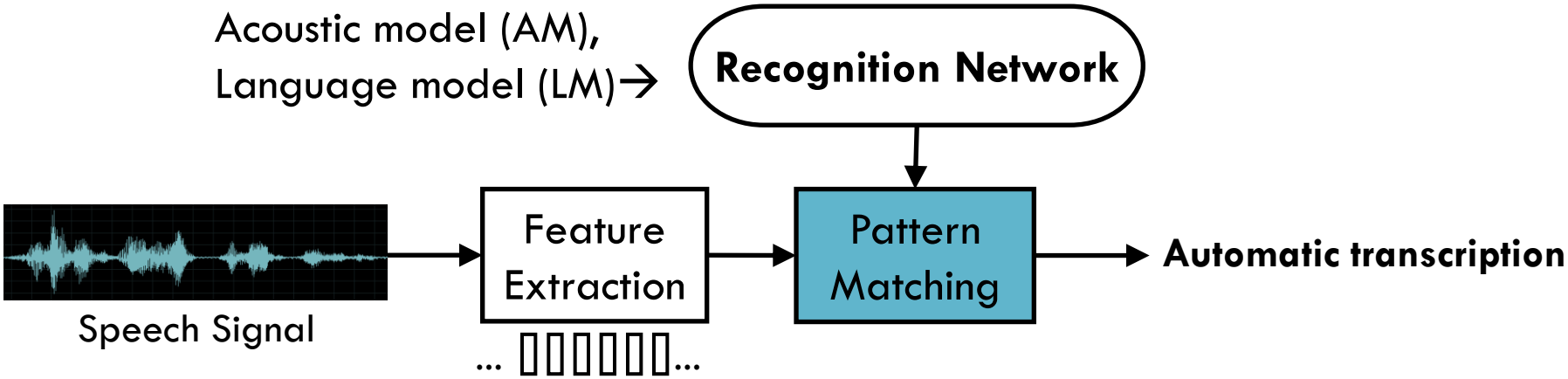
{tarjanb, sarosi, mihajlik}@tmit.bme.hu

tfegyo@aitia.ai

Motivation

- Call centers
  - Important interface between client and company
  - Feedbacks, Opinions, Complaints
- This invaluable information source is usually **not processed**
- Our aim
  - Real-time speech transcription
  - For monitoring call center conversations

Automatic Speech Transcription



- Challenge in transcription of call center speech
  - Spontaneous, conversational style
  - Background, transmission noises
  - Meaningful, speech-like content (**non-verbal events**)
    - Hesitations (<ee>, <*mm*>, <*aa*>), Constent (<*mhm*>)

Objectives

- **Real-time** call center speech transcription
- Baseline system
  - In-domain training data
  - Word-based LM
- Improving baseline system
  - Reduction of data sparseness
    - Additional training texts
    - Morphological language modeling
  - Refined modeling
    - Modeling of non-verbal events

Acoustic training data

- Acoustic training data
  - In-domain data
    - Task-specific recordings (Bank call center)
    - Manually transcribed (regular speech + non-verbals)
    - 25 hours
  - Related domain data
    - Other call center recordings (Insurance company)
    - Manually transcribed
      (regular speech + non-verbals + non-speech noises)
    - 38 hours

Acoustic test data

- Bank call center recordings
  - 2 hours
  - 8 kHz, 16 bit
- 20 recordings
  - 10 incoming calls
  - 10 outgoing calls
- Coding
  - Originally ADPCM
  - Converted to 8kHz/16bit PCM

Acoustic model training

- Three-state, left-to-right Hidden Markov Models
  - Hidden Markov Model Toolkit (HTK)
- Baseline noise modeling
  - Bank + Insur (25h + 38h)
  - Non-speech noise model
- Explicit non-verbal event modeling
  - To avoid confusion with regular word phones
  - Baseline noise modeling
  - + Models for **non-verbal** events
    - Hesitations (<ee>, <*mm*>, <*aa*>), Constent (<*mhm*>)

Training texts

- Manual transcriptions (*In-domain, Related domain*)
- Additional text corpora
  - Bank customer e-mails
  - Bank call center agent training manual

	Manual transcriptions		Additional text corpora	
	In-domain	Related domain	E-mails	Agent manual
Vocabulary	113k	166k	105k	127k
Word forms	10k	21k	19k	2k
Word PPL	131	312	557	267
OOV rate	12%	14%	18%	30%

Word & Morph-based LMs

- SRI Language Modeling Toolkit (SRILM)
- Word-based LMs
  - 3-gram models
  - Exp: *hát megbeszélem a nejemmel*

  /well I talk it over with my wife/
- Morph-based LMs
  - 4-gram models
  - Morfessor Baseline
    - Statistical, unsupervised algorithm
  - Exp: *hát meg beszél em a nejem mel*

Word-boundary reconstruction

- We need words in the output!
- Word boundary "morphs" (WB morph)
  - Exp: *hát # meg beszél em # a # nejem mel*
- Non-initial tags
  - Exp: *hát meg -beszél -em a nejem -mel*
- Modeling non-verbals
  - Exp:

  <mhm> # hát # <ee> # meg beszél em # a # nejem mel

  <mhm> hát <ee> meg -beszél -em a nejem -mel

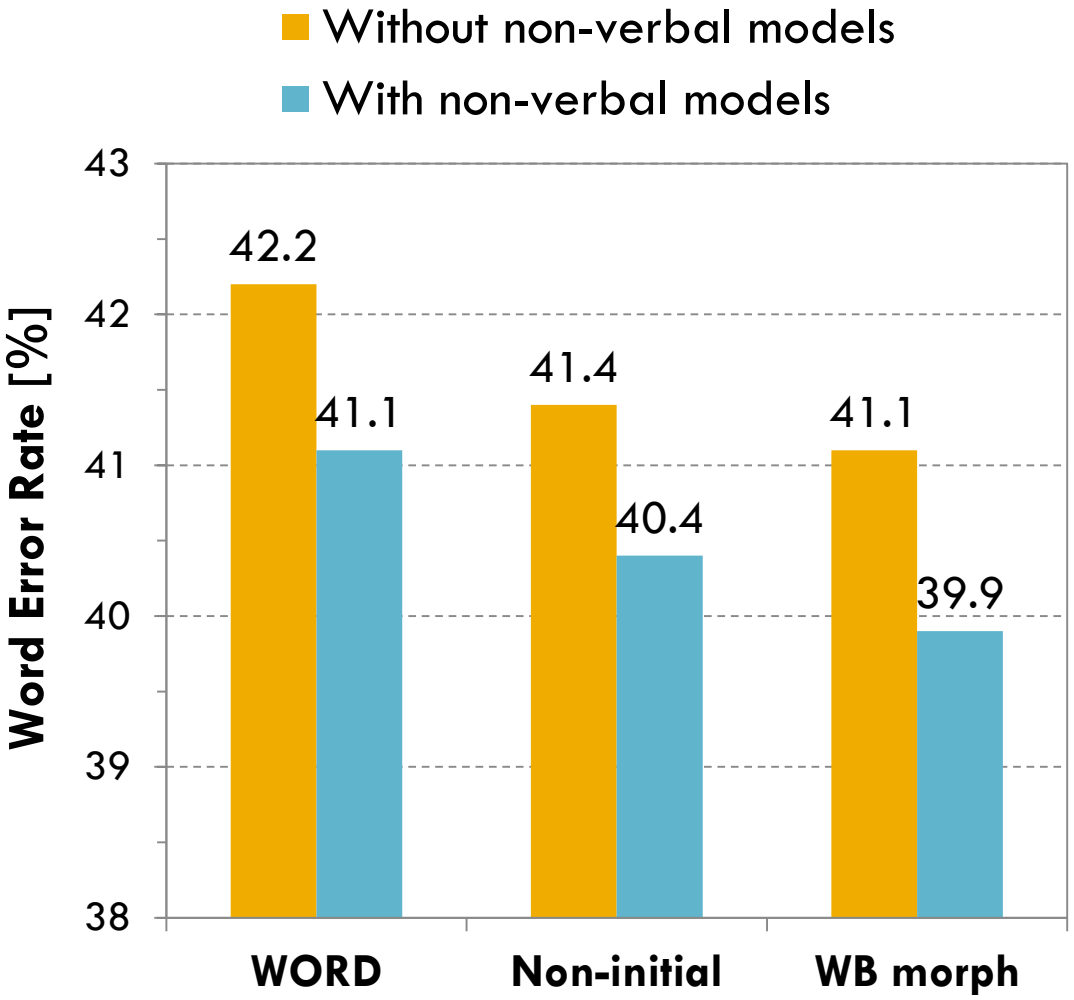
Inclusion of additional LMs

- Baseline system
  - In-domain transcriptions (113k words)
- Extended LMs
  - Linear interpolation
    - Bank + Related domain transcriptions (113k + 166k words)
    - Transcriptions + E-mails (113k + 166k + 105k words)
    - Transcriptions + E-mails + Agent manuals (113k + 166k + 105k + 127k words)
- Interpolation weights of final LM
  - 0.7 Bank trans. + 0.2 Rel. domain trans. + 0.05 E-mails + 0.05 Agent manuals

Network Building & Decoding

- Feature extraction
  - Standard Mel Frequency Cepstral Coefficients (MFCC)
- Knowledge source integration
  - Weighted Finite State Transducer (WFST) networks
  - No vocabulary cutoff!
- Decoding
  - VOXerver - WFST-based decoder
  - Real time operation (2x faster than real time)

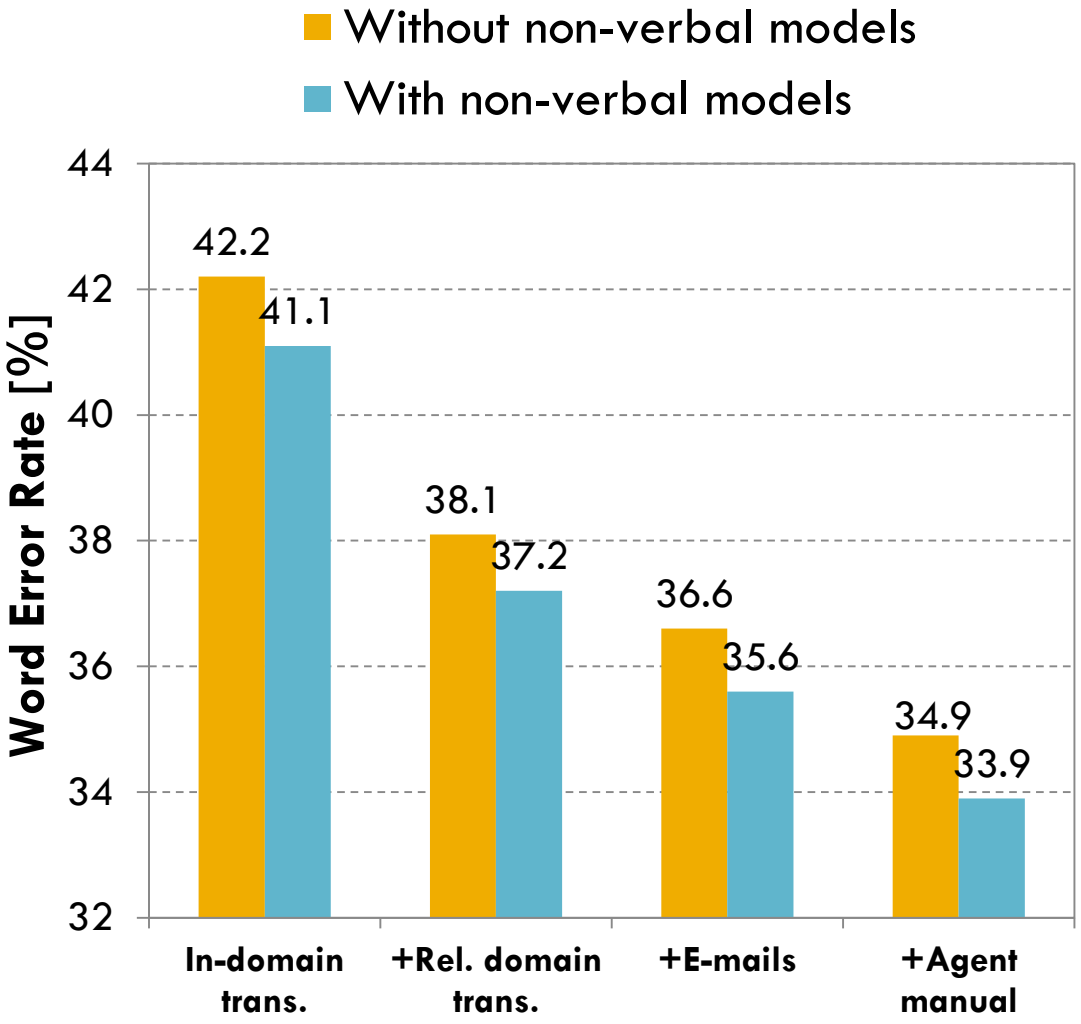
Comparison of lexical modeling approaches



In-domain transcriptions

- Non-verbal models
  - ~3% rel. WER reduction
- Morph-based LMs
  - ~2% rel. WER reduction
- Alltogether
  - From 42.2% to 39.9%
  - ~5% rel. WER reduction
  - Due to less substitution

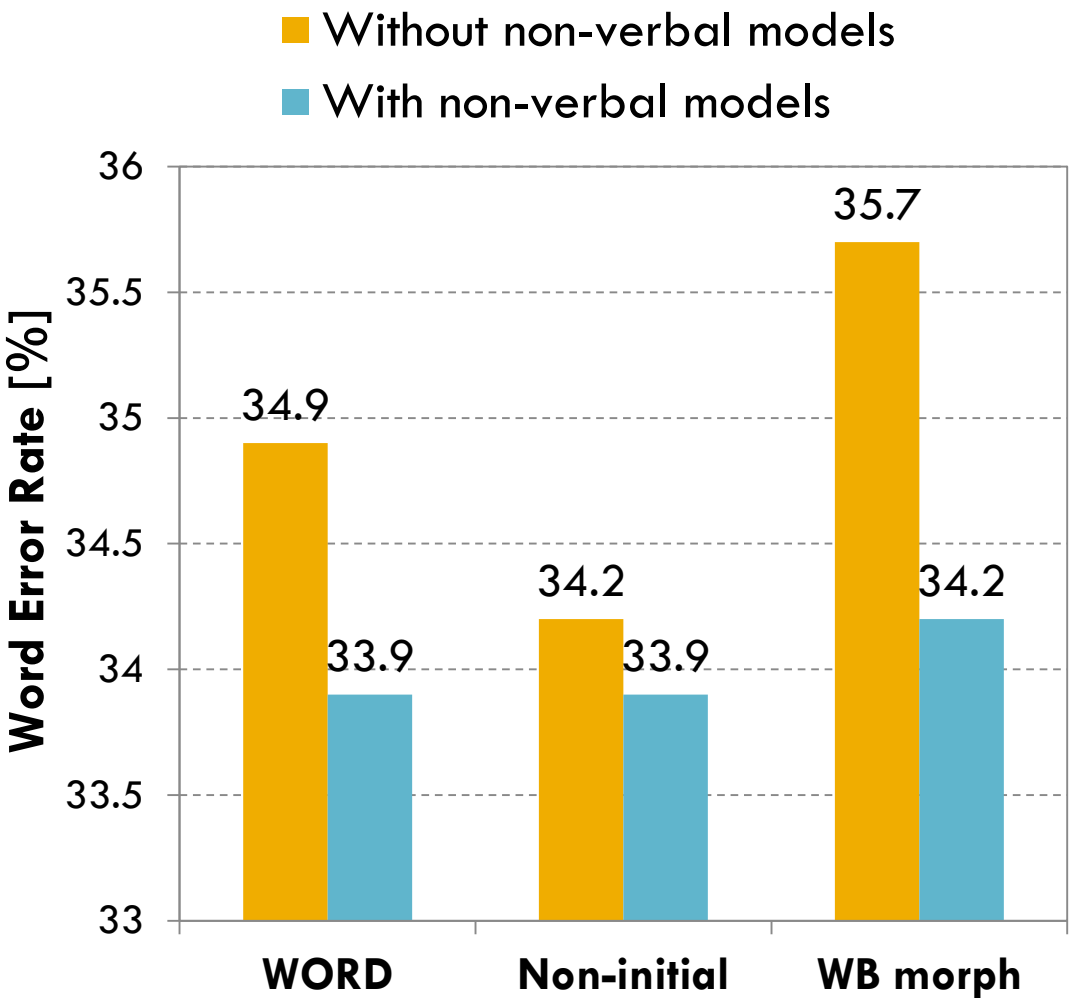
Effect of LM extension



Word-based LMs

- Additional text corps
  - From 42.2% to 34.9%
  - 17% rel. WER reduction
- Non-verbal models
  - ~1% abs. WER reduction
- Alltogether
  - From 42.2% to 33.9%
  - 20% rel. WER reduction

Comparison of lexical models with extended LM



Extended LM

- Non-initial morph LM
  - 2% rel. WER reduction
- WB morph LM
  - 2% rel. WER increase
  - Due to LM interpolation
- **Non-verbal models**
  - Reduce difference among the models

Conclusions

- Improvement of call center transcription system
- Challenging task (real-time trans. of conversational speech)
- Additional training corpora
  - ~17% relative WER reduction
- Morph-based LMs
  - ~2% rel. WER reduction (Non-initial morphs)
- Non-verbal event modeling
  - ~3% rel. WER reduction
- Built for a Hungarian commercial bank
  - Real-life system

Thank you for your attention!

Comments, questions are welcomed!

tarjanb@tmit.bme.hu

Acknowledgement

Our research was partially funded by the KMOP-1.1.1-07/1-2008-0034 (Knowledge Center), KMOP-1.1.3-08/A-2009-0006 (Mindroom), TÁMOP-4.2.2.C-11/1/KONV-2012-0013 (FuturICT.hu), KMR_12-1-2012-0207 (DIANA), AAL-08-1-2011-0001 (PAELIFE) projects.

Nemzeti Fejlesztési Ügynökség

ÚJ SZÉCHENYI TERV

MAGYARORSZÁG MEGÚJUL

A projektek az Európai Unió támogatásával valósulnak meg.