Speech Recognition Results for Voice-controlled Assistive Applications

Alexandru Caranica[1], Horia Cucu[1], Corneliu Burileanu[1], François Portet[2], Michel Vacher[2]

[1]Speech & Dialogue Research Laboratory, University POLITEHNICA of Bucharest

[2] Laboratoire d'Informatique de Grenoble, GETALP Team

Grenoble, France

Presentation outline

- Introduction
- Related Work
- Experimental Setup
- Training and Evaluation Resources
- Language and Acoustic Models
- Experimental Results
- Conclusions & Future Work

Introduction

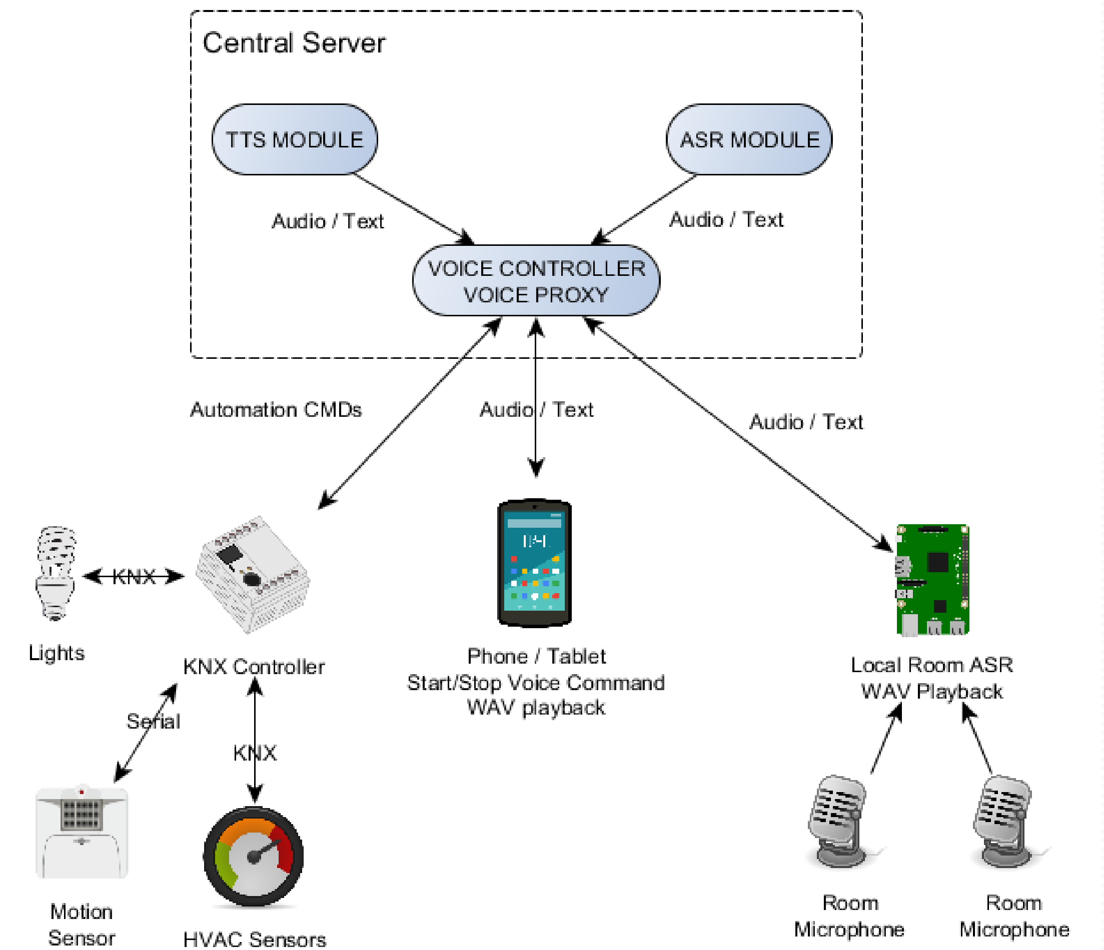
- Recent advances in speech recognition => made voice-controlled smart homes attainable.
- Many companies and communities are providing interfaces or home boxes to make voice control available => Google Home, Amazon Echo, Apple HomePod.
- Most lack customization ability => interoperability with appliances or custom usage scenario is not guaranteed.
- Great performance for widely used languages => little to no efforts were made for under-resourced languages (such as Romanian).

Speech & Dialogue Research Laboratory

Related Work

- Many previous studies focused on protocols and network types => later on putting an accent on user interfaces.
- classical input methods (terminals, touchscreen panels, remote controlled systems) reached maturity => we expect voice control to play a central role in the development of Internet of Things (IoT).
- many challenges need to be solved => robustness against noise, distant speech recognition, the accuracy of keyword spotting and language dependence.
- several projects and joint initiatives => ANVSIB [1], CHIL [2], AMI [3], REVERB [4], PASCAL-CHIME [5], GRID [6] and Sweet-Home [7].

Experimental Setup



Proposed architecture of the intelligent voice controlled automation system.

Training and Evaluation Resources (1)

- Low resource languages => a challenge is to acquire proper speech and language resources, in order to obtain good ASR results.
- Significant efforts were made by our group (SpeeD) to extend the size of the training read speech corpus / to create a spontaneous speech corpus.
- For distant speech recognition experiments => an evaluation corpus was acquired in the DOMUS smart home, from Laboratoire d'Informatique de Grenoble (LIG, http://www.liglab.fr/).
- The evaluation corpus was acquired in realistic conditions => a certain degree of variability has been also taken into account, managing the commands list in such a way that it covers different ways of expressing the same command.

Speech & Dialogue Research Laboratory

Training and Evaluation Resources (2)

- The RCS-train corpus (Read Speech Corpus) => obtained by recording various predefined texts representing news articles and literature, using our online recording application (http://speed.pub.ro/speech-recorder/).
- The SSC-train corpus (Spontaneous Speech Corpus) => created using a lightly-supervised acoustic modeling technique. The originally loosely-transcribed speech data comprised of broadcast conversational speech.
- The SSC2-train corpus => additional spontaneous and read speech, acquired over the Internet. Contains broadcast news, conversational speech, from various Romanian media groups. Segmented and diarized to filter-out all the non-speech parts of the corpus and to create single-speaker utterances.
- The CCA1-eval-cmd corpus (Command Corpus for automation) => contains a limited set of command, recorded with our online application.
- The CCA2-eval-cmd corpus => acquired in realistic conditions (DOMUS lab), to fit a certain set of commands that a user can give to a smart home's devices and utilities. Contains all possible utilities that a user can operate throughout the central controller.

Training and Evaluation Resources (3)

Corpus name	Type of speech	Size	#Speakers
RCS-train	read speech	106 hrs	179
SSC-train	conversational speech	31 hrs	unknown
SSC2-train	conversational + read speech	100 hrs	unknown
CCA1-eval-cmd	read speech	1 hrs	5
CCA2-eval-cmd	conversational + read speech	1 hrs	11
CCA2-eval-nocmd	conversational + read speech	½ hrs	11

Training and Evaluation Resources (4)

- We compared two different features types, for our speech representation:
  - Baseline models => traditional MFCC speech features plus temporal derivatives (13 MFCC + Δ + ΔΔ).
  - Noisy models => noise robust features introduced (Power Normalized Cepstral Coefficients - PNCCs).
- Challenging problem => recognition accuracy degrades significantly if the test environment is different from the training environment, or if the acoustical environment includes disturbances such as additive noise, channel distortion, speaker differences, reverberation, etc.
- PNCC features => bring the most important gains in accuracy in noisy environments.

Speech & Dialogue Research Laboratory

Language and Acoustic Models (1)

- Main ASR system => CMU-Sphinx 4 speech recognition toolkit (core of Large-vocabulary continuous speech recognition – LVCSR - engine).
- Local room end-nodes => run on PochetSphinx, lightweight variant of the Sphinx speech recognition engine.
- A set of preliminary evaluation of the performance of our system was also run on Kaldi => added support for feature-space discriminative training and deep neural networks.
- Acoustic models => 5-state HMMs with output probabilities modeled with GMMs.
- 36 phonemes in Romanian were modeled as context dependent phonemes, with 4000 HMM senones.

Language and Acoustic Models (2)

Name	Training Corpus	Speech Features	# HMM Senones	# GMs
AM001	RSC-train + SSC-train (137 hours)	PNCCs + Δ + ΔΔ	4000	128
AM002	RSC-train + SSC-train (137 hours)	MFCCs + Δ + ΔΔ	4000	128
AM010	RSC-train (106 hours)	PNCCs + Δ + ΔΔ	4000	64
AM054	RSC-train (106 hours)	MFCCs + Δ + ΔΔ	4000	64
AM053	RSC-train + SSC-train + SSC2-train (237 hours)	PNCCs + Δ + ΔΔ	4000	128
AM055	RSC-train + SSC-train + SSC2-train (237 hours)	MFCCs + Δ + ΔΔ	4000	128

Features / Acoustic model descriptions

Language and Acoustic Models (3)

- A Finite State Grammar (FSG) was used for the commands database => split into five categories (exterior, security, multimedia, hvac and electric).

```
……
<electric> = <lumini> | <culoare_lumina> | <dimmer> | <alimentare> | <panouri_solare>;
<lumini>=(aprinde|stinge|opreşte|porneşte|închide|deschide)(lumina|[toate]
luminile)[[de] la parter|din casă];
<culoare_lumina>=(schimbă|aprinde|porneşte)(culoarea luminii|lumina)(roşie|înroşu|[în]
verde|[în] albastru|[în] albastră|[la]normală|la normal);
<dimmer>=(măreşte|creşte|micşorează|scade) luminozitatea [până] la <numar_zeci> la sută;
<alimentare>=(porneşte|reia|opreşte)(alimentarea|încărcarea bateriei);
<panouri_solare>= care este starea bateriei|este încărcată bateria;
//-----------------------------------------------------------------------
  public<comandaCasa>= casandra (<exterior>|<securitate>|<multimedia>|<hvac>|<electric>);
```

Excerpt of the FSG grammar used for commands

Experimental Results (1)

- New features type (PNCC) => tune decoding parameters => obtain optimum results (best SER – Sentence Error Rate):
  - Probability of transitioning into the phone-loop => Out-of-Grammar Probability – OOGP.
  - Relative threshold used in path pruning => Relative Beam Width – RBW.
  - Language Weight (LW) / Word Insertion Penalty (WIP) => kept on default values.

Name	OOGP (1E-)	WIP (1E-)	LW	RBW (1E-)
AM001	40	10	2	70
AM010	30	10	2	70
AM053	40	10	2	70

Best Performing Parameter Values For PNCC Features

Experimental Results (2)

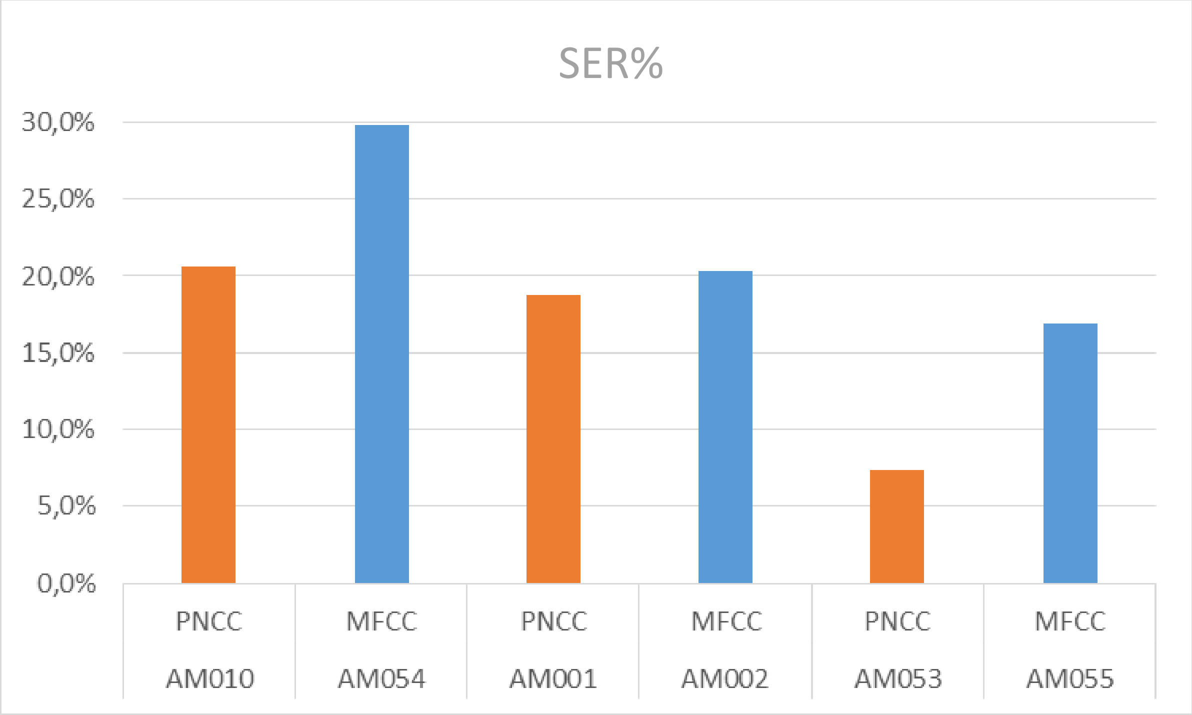
The effect of using Noise Robust Features / Corpus size

The relative SER reduction (CCA1-eval-cmd corpus, recorded in our lab, clean speech):

- 2% for AM001
- 16% for AM010
- 0% for AM053

Experimental Results (3)

The effect of using Noise Robust Features / Corpus size

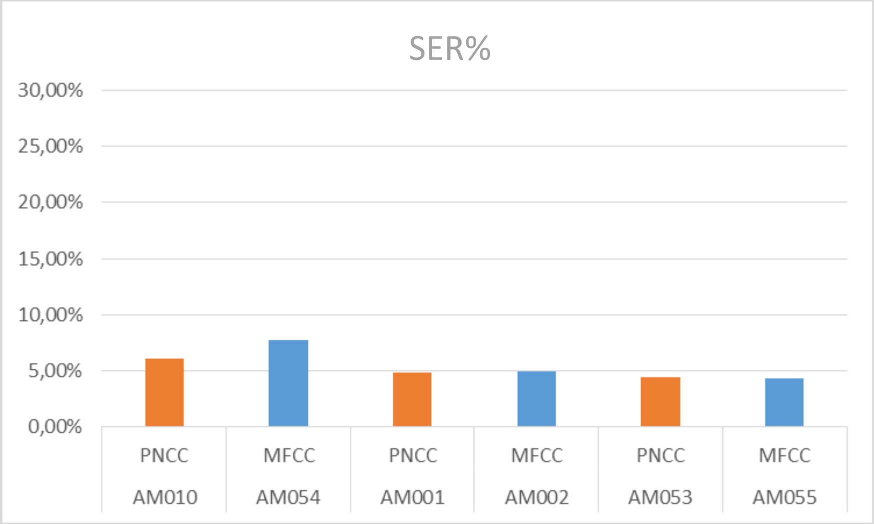


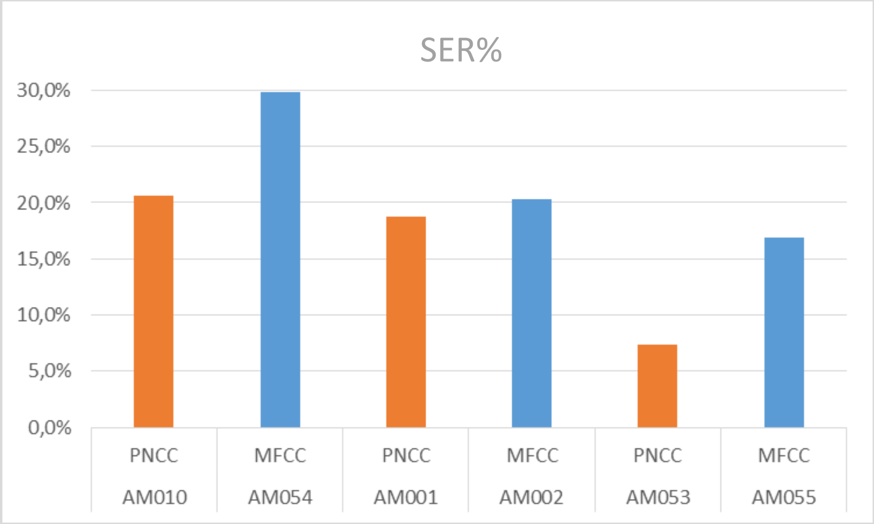
The relative SER reduction (CCA2-eval-cmd, recorded in DOMUS lab, real-life conditions):

- 7% for AM001
- 30% for AM010
- 55% for AM053

Experimental Results (4)

The effect of speaker numbers and type of speech

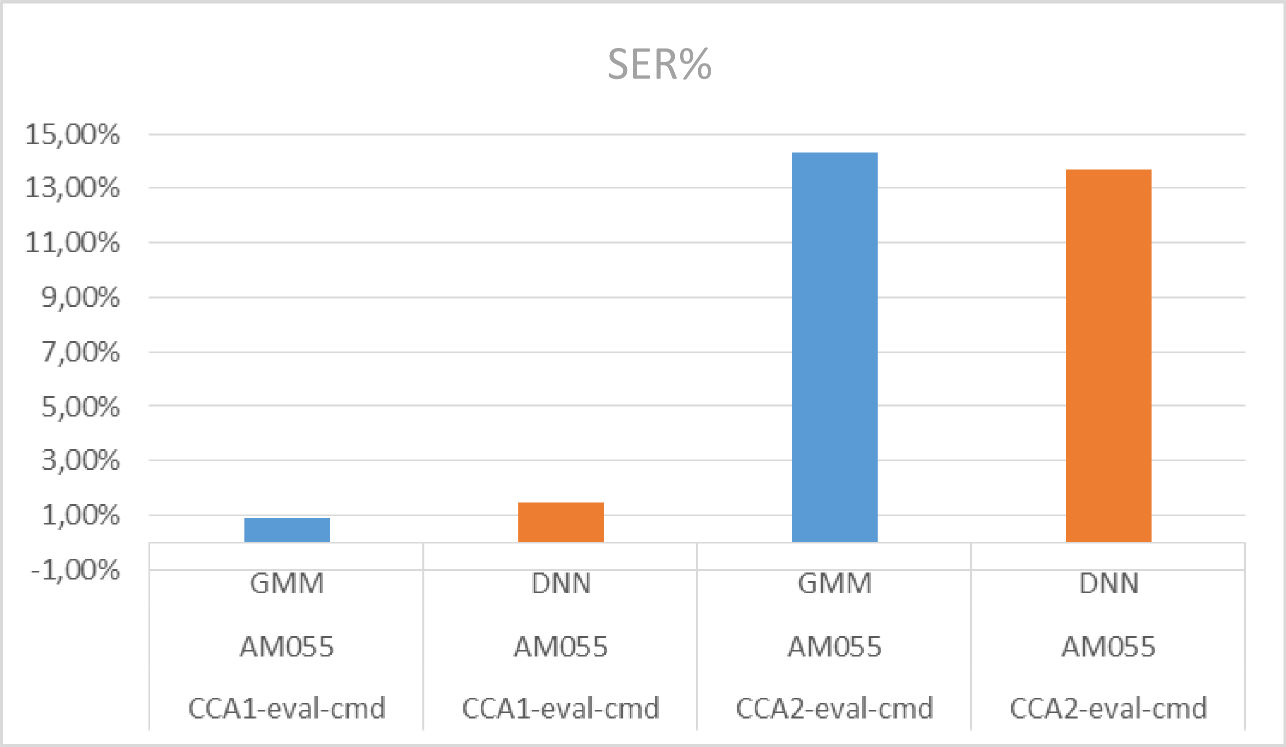




CCA1-eval (clean speech) > CCA2-eval (noisy real-life conditions – DOMUS lab)

Experimental Results (5)

Preliminary results using a DNN toolkit (Kaldi)



Mostly the same results => test on larger datasets => should discriminate better

Experimental Results (6)

Evaluation results on speech without commands

Mixed results => should be analyzed further

All models performed well enough for the given task!

Conclusions & Future Work

- Presented a set of experiments => building a series of acoustic / grammar models for Romanian language, for DSR environment.
- Used in-house acquired corpora => recorded in real-life conditions => better predict the performance of our models.
- Our best PNCC model => relative improvement of up to 55%, compared with the same MFCC acoustic model.
- Future work => proposed to cover multilinguality.
- Future work => migration towards a DNN toolkit (Kaldi) => promising results.

References

[1] Speech & Dialogue Research Laboratory, University Politehnica of Bucharest, "Natural-language, Voice-controlled Assistive System for Intelligent Buildings (ANVSIB)", http://speed.pub.ro/anvsib, project ID: PN-II-PT-PCCA-2013-4-0789, contract number 32/2014.

[2] D. Mostefa et al., "The CHIL Audiovisual Corpus for Lecture and Meeting Analysis inside Smart Rooms," Language resources andevaluation, vol. 41, no. 3, pp. 389–407, 01/2008 2007.

[3] J. Carletta et al., "The AMI Meeting Corpus: A Pre-announcement," Proc. of the Second International Conference on Machine Learning forMultimodal Interaction, 2006.

[4] K. Kinoshita et al., "The REVERB challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech," in IEEE Workshop on Applications of SignalProcessing to Audio and Acoustics (WASPAA), 2013, pp. 1–4.

[5] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge.,"Computer Speech and Language, vol. 27, no. 3, pp. 621–633, 2013.

[6] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," The Journal of the Acoustical Society of America, vol. 120, no. 5, pp. 2421–2424, November 2006.

[7] M. Vacher, S. Caffiau, F. Portet, B. Meillon, C. Roux, E. Elias, B. Lecouteux,, P. Chahuara, "Evaluation of a Context-Aware Voice Interface for Ambient Assisted Living: Qualitative User Study vs. Quantitative System Evaluation", in ACM Transactions on Accessible Computing (TACCESS) - Special Issue on Speech and Language Processing for AT, vol. 7 (issue 2), pp.5:1-5:36, 2015.

Thank you!