Comparison of Multiple System Combination Techniques for Keyword Spotting

William Hartmann, Le Zhang, Kerri Barnes, Roger Hsiao, Stavros Tsakalidis, and Richard Schwartz

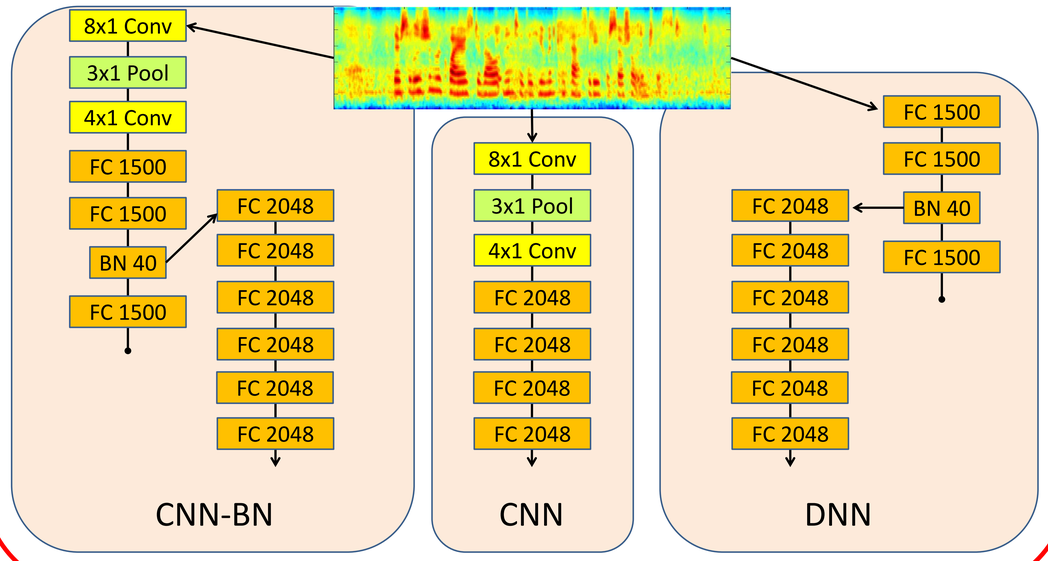
Raytheon BBN Technologies, Cambridge, MA, USA {whartman, lzhang, kbarnes, whsiao, stavros, schwartz}@bbn.com

Raytheon BBN Technologies

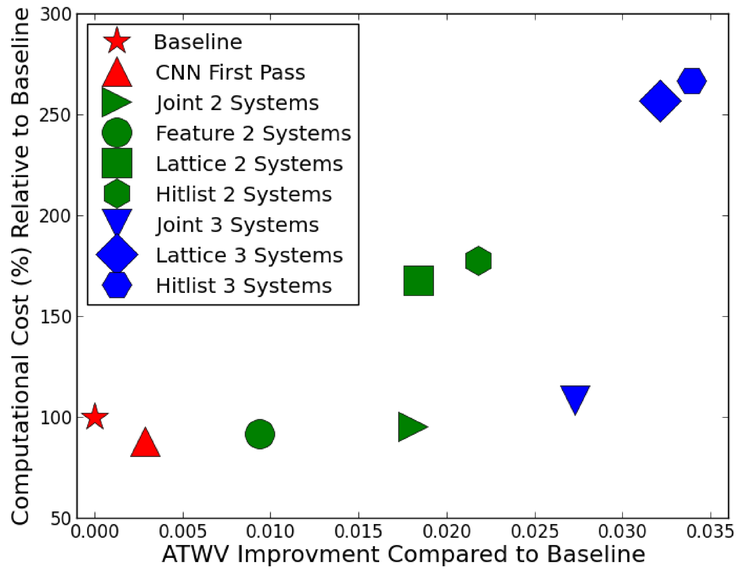
Introduction

- We explore four approaches to system combination: feature combination, joint decoding, lattice combination, and hitlist combination.
- Each approach has its own trade-offs in terms of performance, model restrictions, and computational cost.
- We report results on four languages from the IARPA Babel Program.
- Our focus is on keyword spotting (KWS) and the actual term-weighted value (ATWV) evaluation metric.
- While hitlist combination gives the best performance, lattice combination gives nearly identical performance with less computational cost.
- Joint decoding also significantly improves performance, with little additional computational effort.

Model Types



Computational Cost vs. ATWV



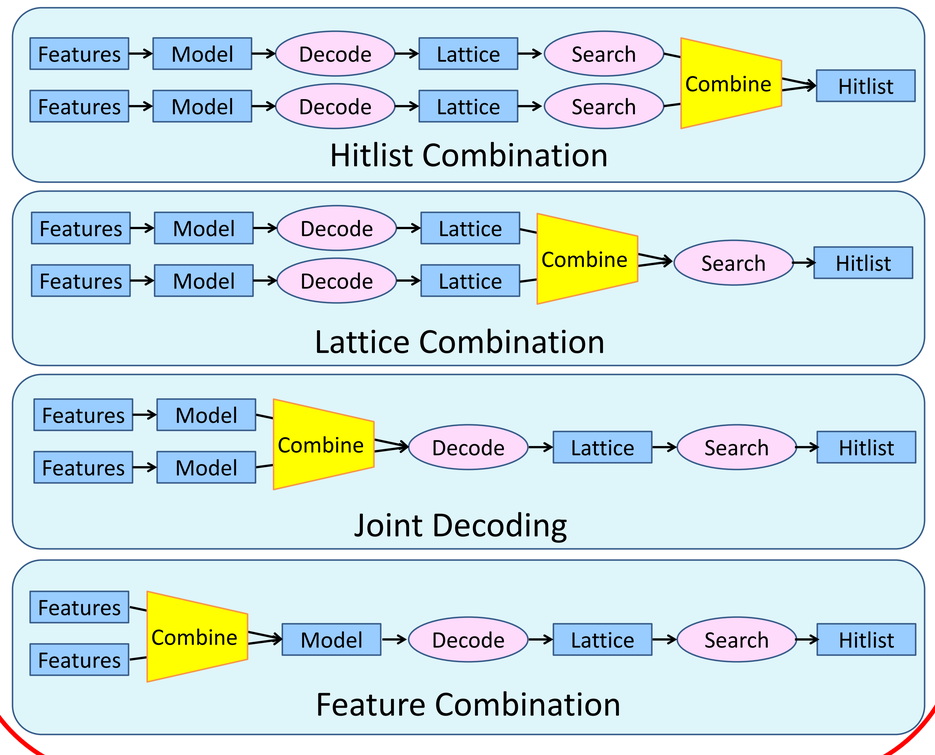
Experimental Setup

- We use the Sage speech recognition toolkit.
- Sage combines BBN's Byblos with open source toolkits such as Kaldi and CNTK.
- Sage also includes a cross-toolkit FST recognizer that supports models built using the various component technologies.
- Three types of models are used:.
  - DNN trained on DNN-based BN features.
  - DNN trained on CNN-based BN features.
  - CNN trained on filterbank features.
- Keyword spotting is performed using both whole word and fuzzy phonetic search.

IARPA Babel Data

- We use four FLP language packs: Amharic (IARPA-babel307bv1.0b), Guarani (IARPA-babel305b-v1.0c), Igbo (IARPAbabel306b-v2.0c), and Pashto (IARPA-babel104b-v0.bY)
- Each language contains about 40 hours of transcribed data.
- Lexicons are built using simple G2P rules.
- Trigram language models are built using only the available transcribed training data.

System Combination Approaches



Results and Conclusions

Language	Baseline	Feature	Joint	Lattice	Hitlist
Amharic	0.583	0.592	0.603	0.606	0.607
Guarani	0.571	0.560	0.582	0.588	0.585
Igbo	0.339	0.351	0.365	0.364	0.365
Pashto	0.411	0.427	0.431	0.437	0.436
Average	**0.476**	**0.483**	**0.495**	**0.499**	**0.498**

ATWV results combining two systems

Language	Baseline	Joint	Lattice	Hitlist
Amharic	0.583	0.606	0.615	0.618
Guarani	0.571	0.590	0.594	0.594
Igbo	0.339	0.366	0.367	0.372
Pashto	0.411	0.440	0.445	0.444
Average	**0.476**	**0.501**	**0.505**	**0.507**

ATWV results combining three systems

- **Feature Combination**
  - Small, inconsistent gains
  - Fastest approach
  - Requires multiple feature types
- **Joint Decoding**
  - Large gains for little additional cost at decode time.
  - Places some restrictions on the models.
- **Lattice Combination**
  - Performance is nearly identical to hitlist combination.
  - Still requires multiple decodings.
  - Can find multi-word hits not present in either lattice.
- **Hitlist Combination**
  - Best performing technique.
  - Places no requirements on individual systems.
  - Most expensive approach.

- Results are reported for the ATWV evaluation metric, a keyword spotting metric where higher values are better.
- Relative improvement decreases as number of systems increase.
- Lattice and hitlist combination give similar performance, but lattice combination requires less computation and storage.
- Joint decoding gives large improvements with little additional cost.

This research was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0013. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.