Automatic Set-up for Speech Recognition Engines Based on Merit Optimization

Gustavo Hernández Ábrego
Xavier Menéndez-Pidal
Thomas Kemp
Katsuki Minamino
Helmut Lucke

Spoken Language Technology
Sony Electronics, Inc.
San José, CA

SONY

Motivation

What's a good real-life speech recognizer like?

- Low errors
- High operation speed

But usually, there's a speed/accuracy condition:

"for low errors, a price is to be paid in speed"

Recognizer, models, etc. set the speed/accuracy trade-off but it also depends on set-up:

- Beam search
- Insertion penalties
- Etc…

Entertainment robot applications

- Demanding recognition platform.
- Accurate and prompt recognition required.
- Parameters set-up to maximize performance.

Biped robot SDR4X

Aibo ERS-220

SONY

Speech recognition merit

- Tedious approach:

  Several different set-ups tried; results tabulated; expert monitors them and decides what's "best".

- This method:

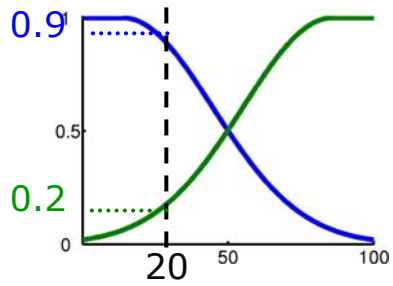
  Recognition set-up is defined automatically through merit optimization. Expert takes "best".

Merit is a combined assessment of recognition performance attributes:

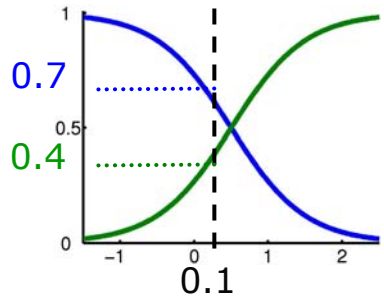
1. Errors.
2. Speed.
3. User-defined trade-off param: WER vs. Time (WoT).

Fuzzy recognition merit

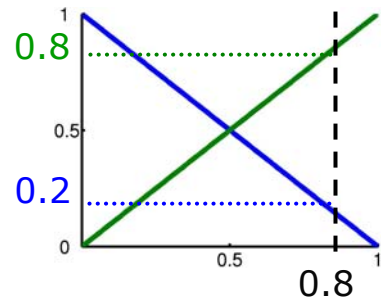
WER



Speed



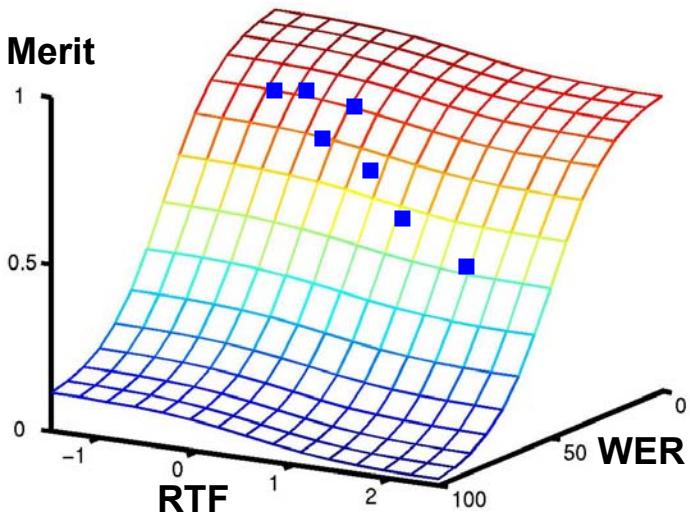
WoT



If WER ↓ and Speed ↓ and WoT ↓, then z ↑
0.9 * 0.7 * 0.2 * 1 = 0.126
If WER ↓ and Speed ↑and WoT ↑, then z ↓
0.9 * 0.7 * 0.8 * 0= 0.000
If WER ↓ and Speed ↑ and WoT ↓, then z ↓ .
If WER ↓ and Speed ↓ and WoT ↑, then z ↓ .
If WER ↑ and Speed ↓ and WoT ↓, then z ↑ .
If WER ↑ and Speed ↓ and WoT ↑, then z ↓ .
If WER ↑ and Speed ↑ and WoT ↓, then z ↓ .
If WER ↑ and Speed ↑ and WoT ↑, then z ↓ .

$$Merit = \frac{\sum\left(z * \prod f(x)\right)}{\sum\left(\prod f(x)\right)}$$

Gradient optimization



WoT=0.9

- Fuzzy merit defines a “simple” cost function.
- Depends on WoT.
- N-space (N=num of set-up params).
- A theoretic maximum.
- Many practical maxima.
- R-Prop optimization.
- Several (10-20) epochs.

Recognition results

Two LVCSR testing applications:

American English Travel Domain

WoT	WER	RTF	epoch	merit
0.80	18.99	0.90	11	0.8813
0.90	16.73	1.08	4	0.9061
manual	17.68	0.99	-	-

Japanese Chat Test

WoT	WER	RTF	epoch	merit
0.80	29.46	0.86	14	0.8053
0.90	28.84	1.21	8	0.8068
manual	29.05	0.92	-	-

Discussion

- WOT effective to handle RTF/WER trade-off.
- Fuzzy merit, based on general patterns, able to handle diverse recognition tasks.
- Derivative: two-point seems inaccurate but its low-cost justifies it.
- Several parameters need rounding.
- Rprop based on gradient direction (not value). Better for integer parameters.
- Parameters heavily correlated. Partial derivatives not independent.

Concluding remarks

- Automatic set-up good to replace tedious manual tuning.
- Recognition results might rise because new regions of the set-up space are explored.
- Methodical procedure avoids human-induced errors and reduces expert input.
- Fuzzy merit suitable as optimization function.
- Fuzzy system robust to application (even language changes).
- Method could be applied to other trade-offs.