Discovering sub-word units with sparse coding

Automatic segmentation and clustering of speech using sparse coding and metaheuristic search

Wiehan Agenbag and Thomas Niesler

Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa

1. Introduction

- ASR: currently very resource intensive:
  - Transcribed audio
  - Pronunciation dictionaries
- **Aims** of this study:
  - investigate the feasibility of **unsupervised determination** of sub-word acoustic units (SWU's)
  - eventually: enabling ASR for **under-resourced languages** by inducing **pronunciation dictionaries**
- Our **approach**:
  - modified form of **sparse coding** and dictionary learning

2. Background

- Scale-invariant **convolutional sparse coding**
  - Weighted sum of time-dilated and time-delayed **basis functions**, which act as **SWU's**
  - Reconstruct input speech signal
  - Codes: coefficients representing the weighted sum
- Reconstruction is evaluated in terms of a **cost function** that consists of a **weighted sum** of
  - the number of non-zero coefficients (hence, **sparse** codes) and
  - signal reconstruction error
- Model is constrained to a **non-overlapping** segmentation of the input speech into SWU's.

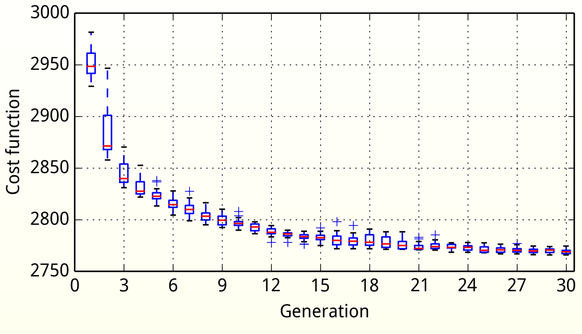
3. Implementation

- **Local search** for SWU's: successive iterative updating of sparse codes and basis functions.
- **Dynamic programming** used to determine optimal solution to sparse coding coefficients.
- Scaled versions of basis functions are updated independently and then aggregated.
- Local search combined with a more global metaheuristic search based on evolutionary approaches.
- **Blind segmentations** form a pool from which SWU's are initialised by random sampling.

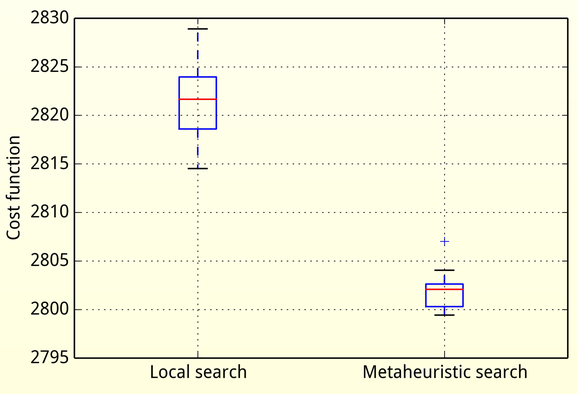
4. Dataset

- **TIMIT** dataset was used for training: 1386 SI utterances from the SI subset of the training partition.
- SA and SX utterances were avoided due to potential bias introduced by repetition.
- Availability of hand-crafted phonetic transcriptions, allows evaluation of discovered SWU's.

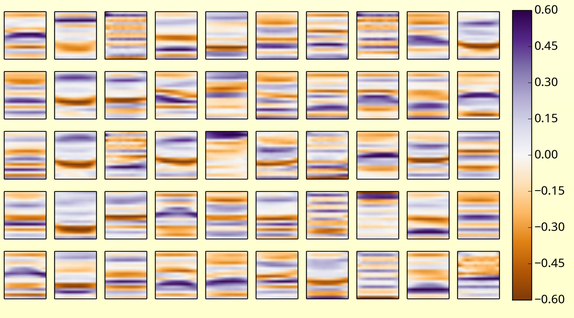
5. Results



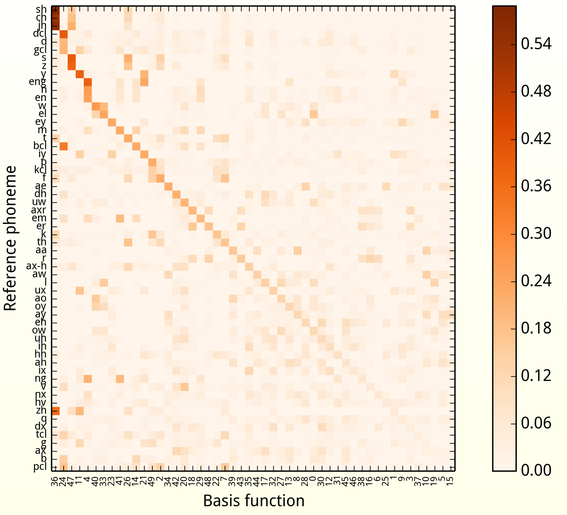
Generational development of the cost function for a population of sets of sub-word units.



Comparison between terminal cost functions for a pure local search and a metaheuristic search. Both searches used the same initialisation.



Spectrogram representation of a set of learned sub-word units.



Coincidence between learned sub-word units and reference phonetic transcriptions. Every row sums to one.

Word	Occurrences	Pronunciations	Top pronunciation	Top 3 pronunciations	Top 5 pronunciations
the	508	55 (22)	14% (39%)	25% (81%)	34% (88%)
a	351	61 (20)	5% (36%)	15% (72%)	22% (81%)
to	269	86 (34)	7% (20%)	18% (43%)	26% (60%)
of	245	82 (25)	9% (35%)	20% (69%)	29% (80%)
and	226	102 (61)	8% (24%)	17% (36%)	24% (48%)
he	212	37 (10)	33% (63%)	51% (91%)	65% (96%)
in	184	75 (19)	8% (43%)	17% (71%)	25% (83%)
is	170	61 (17)	12% (46%)	29% (81%)	38% (88%)
are	92	51 (18)	8% (25%)	20% (57%)	28% (77%)

Fraction of all occurrences of a word that is transcribed by its n-most frequently occurring pronunciations. Statistics for reference transcriptions in parentheses.

6. Conclusions

- Global search for sub-words units show an improvement over a purely local search in terms of cost function.
- Good coincidence of SWU's with reference phonemes.
- Informal listening tests also imply a high quality clustering.
- Extracted word pronunciations show relatively poor consistency, with many different pronunciations.

Presented at: Interspeech 2015, 6 – 10 September, Dresden (Germany)