Experiments with One–Class Classifier as a Predictor of Spectral Discontinuities in Unit Concatenation

Daniel Tihelka Martin Grůber Markéta Jůzová

NTIS – New Technologies for the Information Society,
Department of Cybernetics
Faculty of Applied Sciences, University of West Bohemia, Czech Republic

SPECOM 2016 – International Conference on Speech and Computer

FACULTY
OF APPLIED SCIENCES
UNIVERSITY
OF WEST BOHEMIA

NTIS

DEPARTMENT OF
CYBERNETICS

Unit selection speech synthesis

- still preferred in commercial sphere
- apparently not so attractive for the research (especially contrary to HMM)
- the tricky part is the setting of *target* and *concatenation* costs
- many research papers already published about this topic

Unit selection speech synthesis

- still preferred in commercial sphere
- apparently not so attractive for the research (especially contrary to HMM)
- the tricky part is the setting of *target* and *concatenation* costs
- many research papers already published about this topic

BUT:

- the results are not very consistent (often even in contradiction)!
- costs are hand-tuned

Unit selection speech synthesis

- still preferred in commercial sphere
- apparently not so attractive for the research (especially contrary to HMM)
- the tricky part is the setting of *target* and *concatenation* costs
- many research papers already published about this topic

BUT:

- the results are not very consistent (often even in contradiction)!
- costs are hand-tuned

BUT:

- we have large corpora at disposal in unit selection approach
- which contains plethora of natural–sounding unit transitions
- why not to use them to train AI to decide what is/is not natural

One–class classifier in Unit selection

- we have focused on concatenation cost now
- not new idea, pioneer study published in *One-Class Classification for Spectral Join Cost Calculation in Unit Selection Speech Synthesis* (IEEE Signal Processing Letters, 2010)
  - ➥ based on various distances of MFCC, LPC and spectra,
  - ➥ works surprisingly well (according to authors),
  - ➥ not evaluated in real TTS task
- we aimed at validation of the original results

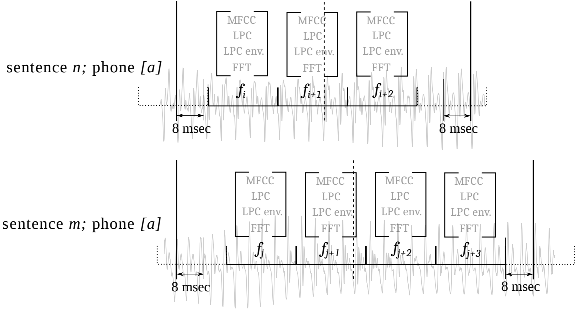
How it works:

- lets take feature distances around natural join
- train OCC on those (continuous) distances
- let the trained OCC to detect if an (unseen) sample is anomaly (unnatural) from its point of view

Features used to train OCC

- OCC is trained to recognise feature distances
- for the experiments we have used:
  - *Euclidean* distance of MFCC,
  - *Mahalanobis* distance of MFCC,
  - *Itakura–Saito* distance of LPC coefficients
  - *Kullback–Leibler* distance of spectral envelopes
- all of these were inspired by the original paper.
- And all the coefficients were obtained from various parametrization
  - **async 20/20** – 20 msec non-overlapped frames (used in the original paper)
  - **async 04/25** – 25 msec frames shifted by 4 msec (provides the most accurate automatic segmentation)
  - **async 12/25** – 25 msec frames shifted by 12 msec (compromise of the previous).
  - **psync pm/25** – 25 msec frames centred around pitch–marks

async 20/20 parametrization:



The distances are computed between neighbouring frames:

- f_i, f_{i+1} / f_{i+1}, f_{i+2}
- f_j, f_{j+1} / f_{j+1}, f_{j+2} / f_{j+2}, f_{j+3}
- but frames too close to the phone boundary were excluded

Listening test

- the (continuous) distances can now be used to train the OCC
- they can also be used to evaluate them (with ACC $\approx$ 99%)

Listening test

- the (continuous) distances can now be used to train the OCC
- they can also be used to evaluate them (with ACC $\approx$ 99%)
- but we want to know how we are good at recognizing non-neighbouring joins
  - ➥ which is how unit selection works

Listening test

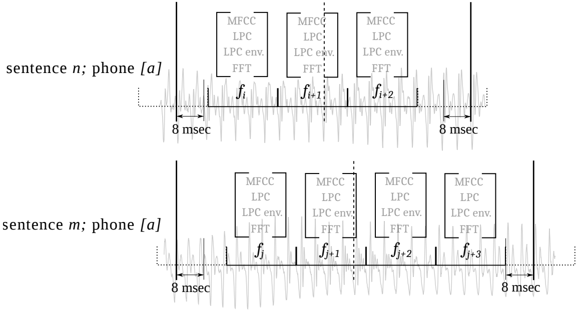
- the (continuous) distances can now be used to train the OCC
- they can also be used to evaluate them (with ACC $\approx$ 99%)
- but we want to know how we are good at recognizing non-neighbouring joins
  - which is how unit selection works
- listening tests were carried out
  - aimed at collecting human perception of natural–sounding/damaged join
  - a set of words was synthesized, randomly combining two halves of a word
  - listeners were asked to evaluate if they hear an unnatural artefact
  - distances of the (non-originally neighbouring) frames were used for evaluation
  - more details in the paper

Listening test

- the (continuous) distances can now be used to train the OCC
- they can also be used to evaluate them (with ACC $\approx$ 99%)
- but we want to know how we are good at recognizing non-neighbouring joins
  - ➥ which is how unit selection works
- listening tests were carried out
  - ➥ aimed at collecting human perception of natural–sounding/damaged join
  - ➥ a set of words was synthesized, randomly combining two halves of a word
  - ➥ listeners were asked to evaluate if they hear an unnatural artefact
  - ➥ distances of the (non-originally neighbouring) frames were used for evaluation
  - ➥ more details in the paper
- still, there are much less examples required for evaluation than it would be required for a binary classifier training

Illustration of Features parametrization

async 20/20 parametrization:



Have a word concatenated at *[a]*,

- left part taken from sentence *n*,
- right from sentence *m*;

the distance is computed for f_{i+1}, f_{j+2}

Data used for OCC training/evaluation

Training

- we have focused on Czech vowels only
- each trained independently,
- the number of distances was limited to 4000 (rnd. sel.),
- 80% of the (continuous) distances were used.

Cross-validation

✓ the remaining 20% of all the (continuous) distances,

× 1/2 of distances evaluated as "artefact" (non-continuous).

Evaluation

✓ all the distances evaluated as natural,

× the remaining 50% of the distances evaluated as "artefact"

Note that when natural continuous distances were used, the ACC was close to 99%

Classifiers used

MGD Multivariate Gaussian distribution:

- all the distances modelled together in one go,
- tied through covariance matrix,
- most similar to that used in the original research

OCCSVM One-class SVM:

- maps distances into a high dimensional feature space via a kernel function,
- provides encouraging results in another experiment (Interspeech 2016)

GRT Grubbs' test:

- detect multidimensional distance vector as outlier when any of the individual features is detected outlying

All of them from *SciKit* learn toolkit

Results

Results

- Results are rather shuffled ...

Results

- Results are rather shuffled ...
- phones *[a]* and *[o]*, async 20/20

	OCSVM *[a]*	MGD *[a]*	GRB *[a]*	OCSVM *[o]*	MGD *[o]*	GRB *[o]*
TP	29	60	33	36	49	48
FP	4	9	5	16	20	20
TN	5	0	4	5	1	1
FN	31	0	27	14	1	2

- ➥ MGD cannot detect outliers almost at all (??!!),
- ➥ others are rather bad in it as well,
- ➥ GRT detects *[o]* better than OCSVM, but fails e.g. on *[i]*

Results

- Results are rather shuffled ...
- phones *[a]* and *[o]*, async 20/20

	OCSVM *[a]*	MGD *[a]*	GRB *[a]*	OCSVM *[o]*	MGD *[o]*	GRB *[o]*
TP	29	60	33	36	49	48
FP	4	9	5	16	20	20
TN	5	0	4	5	1	1
FN	31	0	27	14	1	2

  - ➥ MGD cannot detect outliers almost at all (??!!),
  - ➥ others are rather bad in it as well,
  - ➥ GRT detects *[o]* better than OCSVM, but fails e.g. on *[i]*
- further analysis suggests that *continuous* distances can be separated from *natural–sounding* distances rather then *natural–sounding* from *artefact* distances
  - ➥ the given set of features does not capture human cognition

Results

- Results are rather shuffled ...
- phones *[a]* and *[o]*, async 20/20

	OCSVM *[a]*	MGD *[a]*	GRB *[a]*	OCSVM *[o]*	MGD *[o]*	GRB *[o]*
TP	29	60	33	36	49	48
FP	4	9	5	16	20	20
TN	5	0	4	5	1	1
FN	31	0	27	14	1	2

  - ➥ MGD cannot detect outliers almost at all (??!!),
  - ➥ others are rather bad in it as well,
  - ➥ GRT detects *[o]* better than OCSVM, but fails e.g. on *[i]*
- further analysis suggests that *continuous* distances can be separated from *natural–sounding* distances rather then *natural–sounding* from *artefact* distances
  - ➥ the given set of features does not capture human cognition
- to allow results verification, we placed the data to github (address in the paper)
  - ➥ we will continue to do so for the future work

Future work

- Error analysis
  - examine classification failures in details (e.g. are human judgments correct?)
  - are the given features suitable?
- Larger evaluation dataset
  - we have extended the number of evaluated word joins
  - we are extending vowels with small number of samples, building artificial (yet meaningful) words
  - use all the listeners evaluations in a reliable way (ICSP 2008), some evaluations were excluded now
- Feature redefinition
  - about to start experiments with distances based on F_0 and F1–F4 formant freqs.

End of presentation.

Thank you for your attention.