Abstract

GTTS systems were developed for the fixed training condition, following the Total Variability Factor Analysis (i-vector) approach, with either Mel-Frequency Cepstral Coefficients (MFCC) or Phone Log-Likelihood Ratios (PLLR) as features. Different classifiers and scorings were applied on top of the i-vectors, and several combinations of them were fused for the final submissions.

Datasets

- Low-energy sections removed from all the signals provided by NIST
- The resulting signals cut into 30-second speech segments
- Set of segments partitioned into three subsets: training, development and test, as follows:
  - Languages with more than 800 segments: 150 segments selected for development 150 segments for test The remaining segments used for training
  - Languages containing between 300 and 800 segments: 150 segments selected for development The remaining segments used for training
  - The remaining languages (with few data) handled as follows: Cantonese: 100 segments for development, the remaining ones for training British English and Brazilian Portuguese: 30 segments for development, the remaining ones for training
- Segments extracted from a given signal allocated to the same set (either training, development or test)
- Development and test sets balanced according to the speech source (CTS and BN, when available)

MFCC Features

- Computed in frames of 25 ms at intervals of 10 ms
- SDC with 7-2-3-7 configuration: 56-dimensional feature vectors
- Frame-level Speech Activity Detection (SAD) based on BUT decoder for Hungarian, performed by removing feature vectors whose highest posterior was found for the integrated non-phonetic unit

PLLR Features

- Phone Posterior Extraction
- KALDI is used to train a NNet-based acoustic model for English, based only on LDC97S62 (Switchboard-1 Release 2) and the Mississippi State University transcripts provided by NIST
- The acoustic model includes 42 phonetic and 4 non-phonetic units
- The acoustic model is applied to extract frame-level phone posteriors from audio signals

- Given a phone decoder that outputs an N-dimensional vector of phone posteriors at each frame: \( p = (p_1, p_2, ..., p_N) \), such that \( \sum_{i=1}^{N} p_i = 1 \) and \( p_i \in [0, 1] \), for \( i = 1, 2, ..., N \), PLLRs are computed as follows:
  \[
  r_i = \log\left(\frac{p_i}{1 - p_i}\right) = \log(p_i)
  \]
- Non-phonetic units are integrated into a single non-phonetic unit by adding their posteriors
- Frame-level SAD in PLLR systems performed by removing the feature vectors whose highest PLLR value was found for the integrated non-phonetic unit

i-vector configuration

- For each set of features (MFCCs and PLLRs), a gender-independent 1024-mixture GMM was used as UBM, estimated by ML using a subset of swb1_LDC97S62 and swbcell2_LDC2004S07
- Total variability matrix estimated on the same training set
- 500-dimensional i-vectors with length normalization

Classifiers

Generative Gaussian (G)
- Fully Bayesian Generative Gaussian (FBG)
- Logistic Regression (LR)
- Neural Network (NN)

Backends

- Fully Bayesian Generative Gaussian (FBG)
- Discriminative Gaussian (DG)

Fusion

- Linear Logistic Regression
- Fusion parameters estimated on the development subset FoCal toolkit

LLR Computation

Log-Likelihood Ratios (LLR) computed from calibrated and fused scores \( s = [s_1, s_2, ..., s_L] \), as follows:

\[
LLR_i = \log\left(\frac{e^{s_i}}{\sum_{j=1}^{N} e^{s_j}}\right)
\]

where \( i \) is the target language, \( C_i \) is the cluster where the target language \( i \) belongs to and \( N \) is the number of languages in \( C_i \)

Primary System \( (C_{avg} = 0.285) \)

Fusion of four sub-systems:

1. PLLR features + FBG classifier + DG backend
2. PLLR features + LR classifier + DG backend
3. PLLR features + NN classifier + DG backend
4. MFCC features + NN classifier + DG backend

Alternative Systems

Alternative systems consisted of different combinations of sub-systems, from a single sub-system up to 6 sub-systems

No performance improvements with regard to the primary system

Conclusions

- GTTS systems for the fixed-training condition based on state-of-the-art technology with no specific tunings (e.g. 30-second segments were used)
- Fusion was advantageous in development, but did not provide any remarkable improvement in evaluation
- Probably, the limited amount of data available led to overfitting to the conditions seen in development
- The huge performance degradation observed from development to evaluation suggests the existence of a mismatch (speakers, channels) between both datasets
- Extremely poor performance attained for some language clusters (e.g. French): it may be revealing additional (unknown) issues