Abstract

Segmenting the speech signals on the basis of time-frequency analysis is the most natural approach. Boundaries are located in places where energy of some frequency subband rapidly changes. Speech segmentation method which bases on discrete wavelet transform, the resulting power spectrum and its derivatives is presented. This information allows to locate the boundaries of phonemes. A statistical classification method was used to check which features are useful. The efficiency of segmentation was verified on a male speaker taken from a corpus of Polish language.

Experimental Results

The WEKA LogitBoost classifier trains the classifiers on weighted versions of the training samples. It gives higher weights for those which are missclassified. The final classifier is defined to be a linear combination of the classifiers from each stage. Logistic Boost uses an adaptive Newton algorithm to fit an additive multiple logistic regression model.

We tested 7 different sets of features for the same classifier and same test data to check which features are useful. The classification was evaluated using popular precision and recall measure which is presented in tables and by percentage of properly classified instances which are given in text for all cases. Two evaluations are provided for every set of features to help in grading the method because we did not manage to find any other similar system to use as a baseline.

We started with one left and one right context subset of features to describe the surrounding part of signal. We included first and second derivatives and both of them were smoothed. That gives 54 features in total. 64 % of test instances were correctly classified. The more exact results using recall and precision evaluation are presented in Tab. 2. The final measure is f-score presented separately for sets of features describing frames with boundaries and without. The second group is named in tables as phonemes. From practical point of view we are interested in detecting boundaries so the evaluation of classification of these frames is crucial. So for the first set of features the most important grade is f-score 0.45 (Tab. 2).

We managed to slightly improve results by leaving the second derivative unsmoothed. There were no other changes in the set of features and 64 % of test instances were correctly classified like for the previous set of feature but the more exact evaluation presented in Tab. 3 indicates some improvement through higher f-score, namely 0.466.

In the next approach we kept the same number and type of features but subband features were normalised by dividing by the energy. In that way 60 % of test instances were correctly classified with f-score only 0.135 (Tab. 4).

We tried also another normalising approach, by dividing all features by a maximum in a given subband for an analysed utterance. Around 64 % of test instances were correctly classified but f-score is also quite low, namely 0.413 (Tab. 5). Surprisingly, none of normalisation methods improved results. Finally, we experimented with wider left and right context, namely we added more subsets of features for signal around the analysed one. We have got 66 % of test instances correctly classified by including two contexts to the left and two to the right. In that case we had a set of 90 features with a relatively high f-score 0.519 (Tab. 6).

To use wider context, namely three to the left and three to the right, we had to skip second derivative as the number of features was too large to be operated by WEKA. In that way we had a set of 84 features. 70 % of test instances were correctly classified, but recall for boundary frames was very low, just 0.162 which caused f-score only 0.263 (Tab. 7). It means that generally this set of features is not effective.

The three to left and one to right context was also checked. In that experiment we used the second derivatives, so we had 90 features. We received correctness of 70 % but f-score for boundaries was again quite low, only 0.302 (Tab. 8).

Phoneme segmentation

We trained and tested our classification model on a male speaker of a corpus of Polish. Our experiments were conducted on speech files with the sampling frequency 16 kHz. The database contains 365 utterances (single letters, digits, names and simple sentences). The start of a phoneme should be marked by an initially small but rapidly rising power level in one or more of the DWT levels.

1. Normalise a speech signal by dividing by its maximum value.
2. Decompose a signal into six levels of the DWT.
3. Calculate a set of power samples in all frequency sub-bands according to the table to obtain the power representations \( p_k(i) \) of the \( k \)th subband.
4. Calculate a set of features which may include some smoothing operations, finding derivatives, context features and normalisation.
5. Use WEKA LogitBoost to classify boundaries and non-boundaries frames.

Discrete Wavelet Transform

The human hearing system is equipped with frequency processing system in the first step of sound analysis. DWT has some features similar to the principles of the operation of human hearing system. The wavelet transform provides a time-frequency spectrum. The coefficients of series are computed where the value of \( n \)th wavelet function at the \( M \)th resolution level under the time discretisation. Due to the orthogonality of wavelet functions we obtain

\[
s(n) = \sum \phi_M(n, aD)\delta(n) = \sum a_n \phi_M(n)\delta(n)
\]

where \( a_n \) are supports of \( \phi_M(n) \). The coefficients of the lower level are calculated by applying formulae

\[
g_n = a_n \phi_M(n)
\]

where \( a_n \) and \( g_n \) are the constant coefficients which depend on the wavelet. By applying recursively the above formula

\[
DWT(s) = \{d_M, d_{M-1}, \ldots, d_1, c_1\}
\]