The Concept of Task Specific Speech Database for VAD Systems

J. Kacur
Department of telecommunication, FEI STU, Bratislava, Slovakia
E-mail: *kacur@ktl.elf.stuba.sk*

Abstract - *The aim of this article is to present a concept of small size, low cost, and stile flexible speech database especially designed for the purposes of construction and evaluation of voice activity detection systems. As an example of the concept the Slovak speech database for VAD systems is introduced. It covers wide range of SNR ratios that can be automatically adjusted, according to both general and particular applications. It consists of specially selected isolated words with time labeled occurrences which keeps its size small while enabling its automatic handling. The evaluation of the database was performed by 2 VAD algorithms and compared to the SPEECON database (Spanish version). Observed outcomes fully justify its design and encourage its use for the development and evaluation of high performance VAD systems.*

Keywords – *Speech database, VAD systems, speech recognition*

1. INTRODUCTION

The speech detection algorithms can be found in many different signal processing applications, usually as parts of more complex architectures. These algorithms are widely employed, especially in the following areas: speech compression and transmission, speech recognition, speech enhancement, etc. Usually the most common detection methods are based on: energy, zero crossing, spectral changes, cepstrum, pitch detection, fuzzy logic, or any combination of these, etc.; [9]. Despite the wide range of existing solutions, there is yet no universal detection algorithm working reliably in all possible settings.

This is mainly due to the great generality and complexity that is required from any VAD. There are 3 main problems involved. The speech itself presents wide variety of different types of signals in any domain (time, frequency, energy, etc.). Even parts of silence are regular speech events (phones like t, k, p, etc.). Second problem that is even more complicated is the existence of ever present additive background noise. Noises can exhibit all possible attributes that are restricted only by the channel. Situation can get dramatically worse if the noise properties changes over the time in the similar rhythm to a regular speech. The third problem is the convolutional distortion introduced by any channel which can be time varying as well.

Thus any design of VAD system must include: selection of proper speech features increasing the discrimination between speech representations and some relevant noises (probably not all), adjustment of the decision rules and thresholds that are represented by bulk of variables. These can only be set up by using proper training and testing speech samples covering the expected environment. Because of that, speech databases are collected for great variety of languages and environments, some of them are: SPEECHDAT [3], TIMIT [2], SPEECON [1], MOBILDAT, etc. However these are almost exclusively design to fit demands for the process of building and testing of speech recognition systems based on context dependent phonemes. This slightly differs from the needs of explicit VAD systems (VAD that works independently of any ARS) and thus most of them are not very suitable for this purpose. However, there are database for speech syntheses, especially corpus based [10], which of course pursuit different qualities. Therefore in the following the design of small size, but flexible database, that meets the demands for explicit VAD systems, is presented and verified.

2. DESIGN REQUIREMENTS FOR THE DATABASE AIMED AT USE WITH VAD SYSTEMS

As mentioned in the introduction, each database used in the real environment should cover all -wide range of settings. Practically speaking, this is impossible to acquire thus we are confined to some most expected cases only. Even this would lead to a huge amount of data, which has its drawbacks like: time consuming collection of samples, expensive construction, gathering of new samples in the case of adding so far uncovered environment and awkward practical handling.

Therefore a good design of speech database for the usage with explicit VAD algorithms should tackle those drawbacks. This can be achieved by several principles.

First it is advantageous to have speech active segments marked for the case of training and testing. This differs from any classical speech databases for recognition, where the problem of time alignment is iteratively solved by the embedded training using proper speech and non speech events' models that

are concatenated. However this is compensated by the need for a huge amount of data which enables the convergence of the method. On the other hand if the data is limited, then even hand made labeling is feasible and it can be even more reliable than when done automatically, where unexpected errors do happen. We should bear in the mind that marking speech active intervals is much simpler than marking word or even phoneme boundaries, thus it is almost always tractable.

As this database is not intended for speech recognition it is not needed to cover all words from the vocabulary or more frequently sub-words units which must be present in statistically relevant numbers. Thus groups of words or sentences can be selected to address the main weak point of each VAD algorithm, which is the classification right at the boundaries of active segments (front end clipping-FEC and OVER features) [9]. This can bring substantial reduction of the database size.

Next, it should cover wide range not only of expected noises, but SNR ratios as well. The recordings should be separated and effectively marked to provide the VAD designer's a great variety of options for the particular VAD set up and testing. Then each VAD can be precisely tailored using this database or its portion.

Finally, its structure should be flexible enough to encompass new kinds of background noises if needed. Thus, it should support its updating facility so that new noises and environments are automatically added such as no extra "manual and tedious" work is needed. This can be easily done if original recordings are made in the clear environment.

3. SLOVAK SPEECH DATABASE FOR THE USE WITH VAD SYSTEMS

To document and verify the abovementioned conditions for the construction of specialized database a Slovak small-sized database has been designed and collected. In the following its brief description is given and details can be found in [4].

For the flexibility, small size and other reasons, it was decided that the database would consist of isolated words originally recorded in the clear environment. The words were specially selected to contain all major phonemes that exist in the Slovak language (approximately 51), at their beginnings and ends respectively. This was led by the necessity to properly train and tests the weakest point of any VAD system which is the detection at the boundaries. Nevertheless, some phonemes are very similar to each other so these minor differences do not play a significant role for explicit VAD systems. Thus their number was reduced to 36 which can be encompassed by 36 words. Each word was preceded and followed by at least 400ms silence interval so the speech occupancy is about 50% over the whole database. These recording were made in a "clear" environment to enable for a flexible addition of required noises. Word boundaries were hand labeled which in fact were the only two tedious tasks. However, as the words were recorded without any significant background noise it can be done automatically as well.

At the presence there are more than 20 speakers whom each recorded 36 words. So far only second order stationary noises common to households like: mixer, hair dryer, blaring TV and artificial white noise were added in several SNR ratios (0, 3, 7, 10, 13, 17 dB) to the original recordings (their SNR is on average 36dB).

4. SPEECON DATABASE

To make a comparison and reference to some of well established professional databases the SPEECON database (Spanish version) [5] was selected. The reason of the choice is that although the database was primarily designed for speech recognition systems its structure indirectly allows it to be used for training and testing of VAD systems in various environments and in a controlled way as well.

It consists of sentences uttered by several speakers. The difference is that each sentence is recorded with 4 different microphones located in different distances from the speaker. Thus there are 4 versions of the same utterance with 4 environments (SNR). SNR of the first microphone is on average 24.4dB whereas the furthest microphone has about -8.64dB

The recordings were made mainly in the office environment. As the exact transcription of each utterance was available, it was possible to train proper ASR system (context dependent). However, only the "clean" utterances were used for training to match the environment for the further time alignment. Finally it was feasible to perform a forced word alignment of "clean" recordings only, to eliminate adverse conditions and possible mistakes that were significant for other recordings. Even though the ASR system form the implicit group of VAD algorithms (potentially best VAD systems) and was trained and applied to "clean" environment, some obvious mistakes did happen. Nevertheless, the estimated speech occupancy is about 77%. Having these time aligned samples for the first microphone it was possible to align data for the remaining ones. However this was not straightforward as the recordings were of different lengths, SNR values, and the time shift varied from one recording to another recording.

5. DATABASE VERIFICATION TESTS USING THE PROCESS OF DESIGN AND EVALUATION OF VAD SYSTEMS

To make any supportive comments or verification of the proposed concept of database, 2 VAD systems were trained and tested on both databases. Then the important exchange of training and testing environments (databases) was performed and compared to the originals results. The applied VAD algorithms [6], [7] are both based on cepstral matrices that are popular for speech recognition and sound discrimination purposes [8]. One algorithm used a linear classification function applied to modified CM matrices to perform the classification. The other one was based on NN classification of CM matrices, which in certain condition can act as a Bayes classifier [5].

Both VAD systems require training phase to acquire proper estimation of adjustable variables. As the training sets approximately 1% of samples from each database were used, whereas the environments contributed equally. There were 15 speakers used in this portion of SPEECON database each of which produced on average 5 minutes of raw data Unlike the SPEECON database the proposed one contained only 11 speakers and less than 1 minute of raw signal for each.

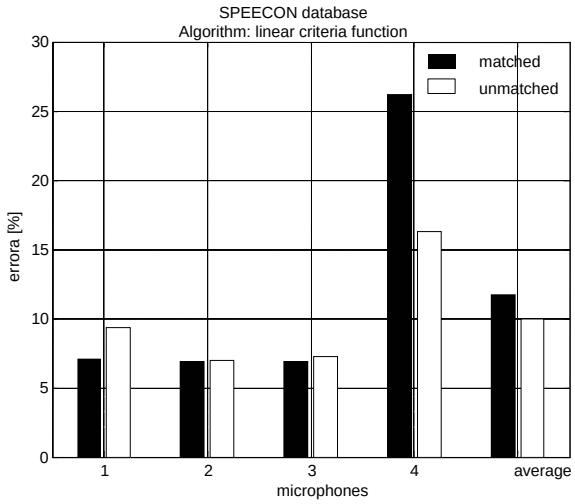

Fig. 1. Speech classification errors using the SPEECON database as the test database. Bars marked as "matched" show the performance of the VAD algorithm train on the same database while "unmatched" bars are for VAD that was trained on the proposed Slovak database. Averaged SNR values of microphones are: 24.4, 12.66, 8.57, and -8.64dB

After all trainings were finished, the testing phase took place over the whole databases. Results are listed in total misclassification ratios separately for all microphones as well as the overall misclassification. In each graph two series are depicted. One marked as "matched" represents the case where the training and testing environments (databases) were the same and the "unmatched" shows the exchanged conditions. Results for the algorithm with linear classification and SPEECON database are shown in fig. 1. The same results executed on the proposed database are in fig. 2. Similar behavior but slightly reduced errors are observed for non-linear classifiers based on NN.

It can be seen that the results in both cases do not exhibits serious discrepancies. Even on the average the training conditions provided by the proposed Slovak database perform better applied to SPEECON database than its original data. This at first glance interesting outcome has its roots in the wider coverage of more adverse environments by the proposed database than in SPEECON database. Inevitably, on the other hand, for higher SNR conditions SPEECON trainings slightly outperform the proposed database, but the effect of preferring low SNR environments seems to be more prevailing on the average.

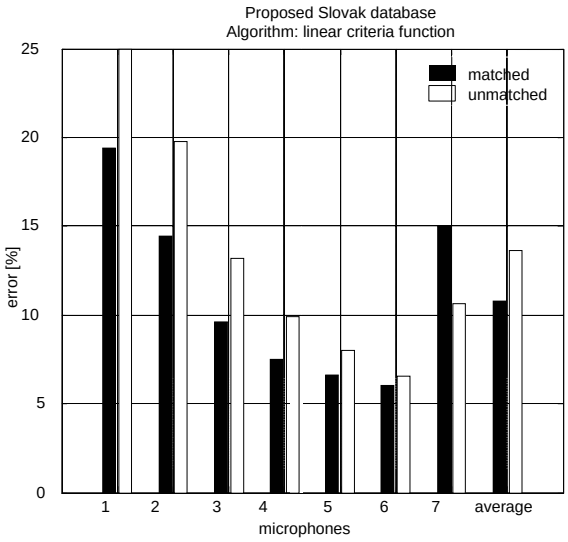

Fig. 2. Speech classification errors using the proposed Slovak database as the test database. Bars marked as "matched" show the performance of the VAD algorithm train on the same database while "unmatched" bars are for VAD that was trained on SPEECON database. Averaged SNR values of microphones are 0, 3, 7, 10, 13 and 17dB

6. CONCLUSION

For the design and evaluation of VAD systems, classical speech databases that are mainly used for recognition purposes, may not provide satisfactory results taking in account their difficult and time consuming collection and huge sizes. Moreover these are usually not very flexible or adaptable to new environments thus often a new database must

be build for changed settings. Here a new concept of speech database especially tuned for VAD system was presented and tested, with the following contributions:

- It significantly reduced the amount of needed data and simplified the process of its building and maintaining while providing comparable or even more robust adjustments for explicit VAD systems.
- Introduced the needed flexibility for adding extra environments by recording the clear set of properly selected words (for the explicit VAD algorithms). Only then the distortion was artificially added in a controlled way using prerecorded real noises that are expected to be relevant in a given application.
- Enabled the classification of utterances according to the SNR values, which makes the database very flexible for VAD designers to better adjust their algorithms.
- Usage of the real small-sized database, designed according to the concept, led to the construction of explicit VAD algorithms performing similarly or in most cases even better, than those algorithms trained on the professional SPEECON database, which is much larger and costly to build.
- As the explicit VAD algorithms should be language independent it is not needed to build it for all languages. Instead the concept of encompassing reach groups of phonemes (statistical occupation may be taken in account) at the boundaries of utterances may be applied. This idea was also well documented by the executed tests where Slovak and Spanish languages were presented and exchanged in the terms of training and testing phases without any deterioration of VAD performance. Instead, the results depended on the training and testing environments in the terms of SNR values. Thus it proofed vital to be able to easily adjust settings in the training phase to match the application environment each time it is needed, which the proposed concept fully supports.

ACKNOWLEDGEMENT

This article was partially supported by the grant "Non-linear processing of multimedia and biomedicine signals in the domain of telecommunications", grant number: VEGA 1/3110/06

REFERENCES

[1] A. Moreno, "Speechdat Spanish Databese for Fixed Telephone Networks". Corpus Design, Technical report, Speechdat Project LE2-4001, 1997

[2] L. F. Lamel, R. H. Kassel and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," in Proc. DARPA Speech Recognition Workshop, Report No. SAIC-86/1546, pp. 100--109, Feb. 1986.

[3] H. Heuvel, V. Galounov, H. Tropf, "The SpeechDat(E) Project : Creating Speech Databases for Eastern European Languages", at: www.fee.vutbr.cz/SPEECHDAT-E/public/conferences/granada98/PAPER_LREC98.DOC

[4] J. Kacur, "Slovak speech database for the design of VAD systems", RTT 2005, Ostrava, Czech Republic

[5] M. D. Richard, R. P. Lippmann, "Neural Network Classifiers Estimate Bayesian a posteriori Probabilities", *Neural Computation*, Volume 3, 1991

[6] J. Kacur, G. Rozinaj, "Word Boundary Detection in Stationary Noises Using Cepstral Matrices", *Journal of Electrical Engineering*, Volume 54, 2003

[7] J. Kacur, G. Rozinaj, S. Herrera, "Speech Signal Detection In The Noisy Environment Using Neural Networks And Cepstral Matrices", *Journal of Electrical Engineering*, Volume 55, 2004

[8] C. Nadeu, D. Macho, "Time and Frequency Filtering of Filter-Bank energies for robust HMM speech recognition", *Speech Communication* 34, Elsevier, 2001

[9] F. Beritelli, S. Casale, "Performance Evaluation and Comparison of ITU-T/ETSI Voice Activity Detectors", *IEEE*, 2001

[10] J. Cepko, M. Turi Nagy, G. Rozinaj, "Low-Level Synthesis of Slovak Speech in S2 Synthesizer". *5th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services*, Smolenice, Slovak Republic. 2005