Segmentation Analysis using Synthetic Speech Signals

M. Greibus, L. Telksnys
Vilnius University Institute of Mathematics and Informatics,
Akademijos St., 4, LT-08663 Vilnius, Lithuania, phone: +370 687 89464
mindaugas.greibus@exigenservices.com

***Abstract*—There is a need of certain number of speech examples to solve real life tasks. A natural speech corpora may suffer from limited phoneme, word or phrase combinations. On the other hand an experimenter can create huge speech corpora using synthetic speech that reflect investigative speech cases that it is similar enough to natural speech. The usage of synthetic and natural speech corpora for speech segmentation algorithm comparison is presented in this paper. The adequateness of synthetic and natural corpora criteria for speech segmentation was proved. The experiments results showed that synthetic signals can be used for speech algorithm research.**

***Index Terms*—Speech synthesis, human voice, adaptive signal detection.**

I. INTRODUCTION

Speech segmentation problem can defined differently depending on application that the segments will be used. Speech segments can be divided into levels: phone, syllable, word and phrase. Automated speech recognition algorithm can solve segmentation tasks indirectly. Blind segmentation is used for task when it is needed extract speech information e.g. speech corpora construction, initial segment extraction to reduce amount of data for ASR. Automatic speech segmentation is attractive for different type of solutions. Cosi et al. [1] define that automated segmentation advantage is that results are predictable: mistakes is done in a coherent way.

In the experiment Wesenick et al. [2] compared manual and automatic phoneme segmentation performances. Results showed that in general errors are done by experts and algorithm for certain phonemes transitions: nasal to nasal and vowel to lateral. Manual and automatic segmentation showed same tendencies for the best segmentation cases also: algorithm and experts were performed well with phoneme transitions: voiceless plosive to nasal, voiceless plosive - vowel.

A natural speech corpus not always provides enough speech cases for experimentation. Also elements in natural speech corpus can suffer from various issues. Some authors [3], [4] in their article propose the use synthetic speech for phone segment alignment. Authors mentioned that synthetic speech does not require having corpora for training, but it is possible generate signals that will be similar enough to natural speech. Other approaches like HMM require large amount of training data and supervised training. Malfrère [4] noticed that phone alignment is working when original signal and synthetic signals are sex dependent.

Sethy and Narayanan [5] were investigating what signal features should be used for aligning synthetic and natural speech signals. Features were tested: Mel-frequency cepstrum coefficients and their deltas, line spectral frequencies, formants, energy and its delta and the zero crossing rate. It was found out that there was no single best feature or feature set for all phonemes. It was proposed to use two feature combinations depending of what phonemes are compared.

Often authors propose to use speech model for phoneme alignment. Synthetic speech can be employed as a part of other speech algorithm also. It can be possible to use speech model for the speech segmentation also.

Manuscript received March 8, 2012; accepted May 22, 2012.

II. ADEQUATE SPEECH MODEL FOR SYNTHESIS

Natural speech signal is a complex process as it is highly dynamic. Speech generation is relaying on multiple mechanisms: linguistic, articulatory, acoustic, and perceptual [6].

Synthesis of speech signal is not a trivial task. Speech signal generation has several steps [7]: text analysis, transforming to computer pronunciation instruction using linguistic analysis and speech waveform generation using recorded speech. Quality of synthetic speech cannot be measured straight forward. Common approach is the qualitative assessment of synthesized speech.

It is challenging task to generate synthetic speech same as natural speech. There are multiple speech engines that can be used to simplify speech signals generation: Mbrola[7], FreeTTS [8] and some others. Mbrola project supports 25 languages. This is very convenient to evaluate multilingual speech algorithms. Mbrola is used by other researches also [3], [5].

Mbrola speech engine generates waveforms from provided phonemes with additional pronunciation information. Text analysis and linguistic analysis should be done separately. It can be done by transformation rules that are used in transformation Letter-To-Sound engines like FreeTTS [8], eSpeak [9]. Such processing use 3 models: phonetisation, duration, pitch. Phonetisation model uses lexical analysis to map graphemes (alphabetic letters, numerical digits, punctuation marks) to phonemes. This mapping converts one stream of orthographical symbols into

symbols of the corresponding sequence of sound. Pitch model [8] determines synthesized speech parameters for pitch, tone, stress, and amplitude. This model is important for naturalness of speech.

In natural speech phonemes are pronounced differently each time. Speech rate of Lithuanian language can fluctuate from 2.6 up to 5.7 syllables per second [10]. In order to imitate duration properties it is possible to define duration as random value $\sim \mathcal{N}(\mu_{DUR}, {\sigma_{DUR}}^2)$. Synthetic speech without randomization property would not represent dynamics of natural speech.

Robustness to noise is another important point when evaluating speech algorithm performance. The stationary noise can be generated by computer and for non-stationary cases real life samples should be used. As synthetic speech signals have no noise it is easy to control noise level SNR measured in dB

$$SNR = 10 \log\left(\frac{\sigma_S^2}{\sigma_N^2}\right), \quad (1)$$

where σ_S^2 – speech variance, σ_N^2 – noise variance.

In order to add noise properly we need to scale speech signal

$$\bar{y}[t] = \frac{\sigma_N \sqrt{10^{\frac{SNR}{10}}}}{\sigma_s} y[t], \quad (2)$$

where $\bar{y}[t]$ – t-th sample of noisy speech signal, $y[t]$ – t-th sample of speech signal.

Synthetic speech can simplify investigation as it is possible to control the noise, content and to generate as many speech utterances as it is needed for an experiment. The model of the speech has to be sufficient to represent natural speech properties.

If model is adequate enough it can be defined by situation-depended criteria [11]: if it is "right", "useful" and "believable". Synthetic speech corpus has to match major requirements for the speech situations that are investigated by researcher. Extreme naturalness may require too much efforts and the effect can be insignificant in comparison with more primitive speech model. Speech model provides ability to control speech content and number of samples. Such tool can help to compare speech algorithms faster than with a specialized natural speech corpus.

The similarity of synthetic speech signal and natural speech can be defined as set of signal attributes that defines signal suitability for communication. In general speech signal quality measurements [13] can be divided in two types: objective and subjective. The objective measurement cannot always unambiguously define quality. The measurement of intelligibility cannot be expressed quantitatively easily and thus are referred as subjective. These methods require expert group and they evaluate how natural sound is.

In this paper segmentation error was used as the quantitative similarity measurement. The first step it is record natural speech signals with selected phrase. Next synthetic signals should be generated using the same phrase. After segmentation algorithms should be applied for both corpora and segmentation errors should be calculated. Similarity of synthetic and natural speech can be expressed as segmentation error difference. If errors values are close, their difference is small thus then synthetic signal is similar to natural speech through the perspective of speech segmentation.

Speech segmentation algorithms require that the speech model should match VIP (Variable, Intelligible, Pronounceable) criteria. Variable – synthesized speech segment properties(duration, amplitude and pitch) must be different each time within limits of natural speech. Intelligible – acoustical signal must be recognized as speech and words must be understood. Pronounceable – phonemes of specific language should be used.

Speech is random signal by nature. Same word will be pronounced each time differently according to context. Ignoring this property speech model would be too static in comparison with the natural speech. As it would be expected speech model should generate signals that it is understandable by human. In general it is not acceptable reuse of different languages phonemes for experiments, as each language has its own specific pronunciation.

III. SPEECH MODEL FOR SEGMENTATION ALGORITHM

Speech segmentation algorithm has to be tested with different speech signals and environments. It is important to control environment during the experiments. This allows indentify defective points easier in comparison with other algorithms and optimize algorithm parameters. Generated signals can cover cases that are rare in natural language. If segment boundaries are known in advance, segmentation errors are eliminated. A speech model, which matches VIP criteria, will be less complex comparing with natural speech, but the signals quality will be understandable by humans.

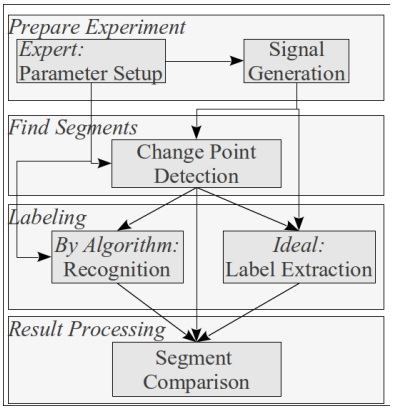

Fig. 1. Speech algorithm analysis using synthetic signals.

Segmentation result evaluation method with synthetic speech is defined in Fig. 1. Synthetic speech corpus parameters must be prepared first of all. The set of phoneme and type of noise have to be specified. The next step is speech synthesis generation. Segmentation algorithm will detect boundaries and label segment. In the last step segmentation results are compared with ideal segments. In this paper corpora for comparison of two segmentation algorithms were used with the phrase: "What is the time?". 50 instances were generated for each value of noise level of white noise: 30dB, 15dB, 10dB, 5dB, 0dB (250 utterances in

total).

The transform of textual information to speech engine instructions are needed next. Simple transformation can be done in two steps: first step it is map one or more graphemes from text alphabet to speech engine understandable symbol and the second step is definition of phoneme duration. Mapping from regular alphabet to computer readable phonetic alphabet depends on speech engine. Mbrola uses SAMPA [14] alphabet.

When speech model parameters are prepared for speech generation we can start generating needed number of speech utterances, which are different through speech segmentation algorithm perspective. Each generated signal should have transcription and audio files. A speech segmentation algorithm processes audio files and retrieves segment sequences. Each retrieved segment is matched to the reference signals and the label is assigned.

Next step is comparison of retrieved segments and original segments. Each segment is matched with reference in original transcription file and correctness of boundaries and labels are evaluated.

IV. SEGMENTATION RESULT EVALUATION AND EXPERIMENT DATA

Fig. 2 shows possible segmentation errors. Original transcription (IDEAL) is compared to retrieved segmentation (AUTO) results. The first task is to verification of segment boundaries. Few error types may be detected (see Fig. 2): EXS – noise was identified as segment; SFT –segment boundary was shifted; ITR – additional boundary was inserted in a segment; JON – joined boundary between two segments:

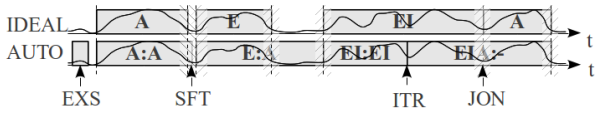

Fig. 2. Automatic speech segment comparison with reference.

Segment boundary detection error rate is calculated

$$ERR_B = \frac{EXS + SFT + ITR + JON}{N_F} 100, \qquad (3)$$

where ERR_B – boundary detection error rate, N_F – number of detected segments.

Non speech and joined segments are not used for labelling. Recognition result can be rejected if pattern matching algorithm cannot distinguish to which class target sample belongs. Recognition error rate 0 is error of the first kind when algorithm says that it was segment class A, but it was segment class E (Fig. 2). It is calculated

$$ERR_R = \frac{AE}{N_B} 100, \qquad (4)$$

where ERR_R – recognition error rate, N_B – number of correct segments from boundary detection step, AE – number of incorrectly labeled segments.

Records of synthetic speech and natural speech were used in order to see whether generated signals can be used for comparison of speech segmentation algorithms. For the experiment was created speech corpora as described above with the phrase: "What is the time?". Two segmentation algorithms were used [15]: Multi Feature Extremum and Rule Based Segmentation (MEaRBS) and the threshold algorithm. It will be shown that both algorithms showed similar results both with synthetic and natural speech, but their quantitative segmentation results allow identifying the better one.

Synthetic speech corpus was created using VIP criteria matching Letter-To-Sound engine and Mbrola with one English speaker voice (us2 male voice). The speaker dependent natural speech corpus was chosen for the experiment. Speaker independent corpus should require larger speech corpus. There were made 51 records of the same phrase pronounced by one male speaker. One utterance was used to train speech recognition engine.

All natural speech utterances where segmented in a semi-automatically way. Segmentation was done in two steps: boundary detection and segment label definition. The results of expert segmentation of natural speech were used to compare with automated segmentation results.

V. EXPERIMENT RESULTS

MEaRBS algorithm was used for segmentation of 300 utterances (250 synthetic and 50 natural speech). Fig. 3 gives segmentation results for all speech types: natural speech, total of all synthetic signals and synthetic speech with various noise levels: 30dB, 15dB, 10dB, 5dB, 0dB. Horizontally it is representing case percent value. We can see that worst results are for synthetic speech with 0 dB noise level. Case of natural speech gives similar result as the case of synthetic speech with 30dB noise level.

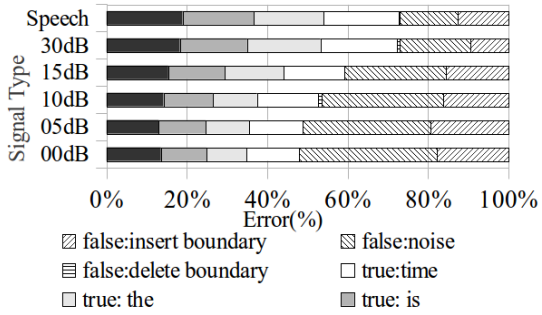

Fig. 3. MEaRBS algorithm boundary detection cases. Speech – represents natural speech represent the results of the experiment.

Error rates are given in Fig. 4. Segment boundaries for natural speech and synthetic speech with 5dB noise showed similar result, but natural speech has higher recognition error rate. Recognition rate of synthetic signal with 10dB noise is similar with natural speech.

Another experiment was done with static threshold segmentation algorithm. The same experimental data and identical system parameters were used. Results are given in Fig. 5. Natural speech segmentation results are similar to synthetic speech with 5dB and 10dB noise level, but segment recognition rates are higher. The reason was the sensitivity of DTW algorithm to segment boundaries.

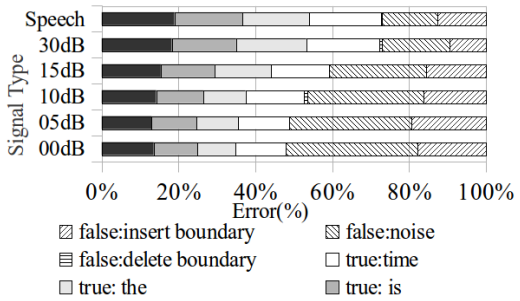

Fig. 4. MEaRBS algorithm segmentation and label class recognition errors.

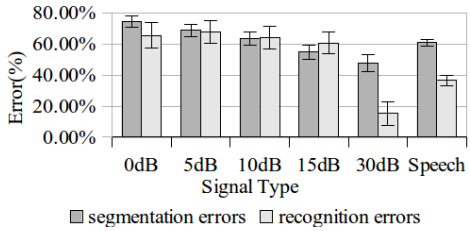

Fig. 5. Threshold algorithm segmentation and label class recognition errors.

VI. CONCLUSIONS

The method of speech segmentation algorithm comparison using synthetic speech corpora was presented. It was proposed segmentation evaluation method with synthetic speech. Such speech signals must match defined criteria. It was shown that results of segmentation results using natural speech and synthetic speech are similar: speech segmentation error is in the confidence intervals of 15 dB synthetic signals. It can be noted that synthetic speech can be used for speech research and some algorithms comparison and investigation of certain cases.

REFERENCES

[1] P. Cosi, D. Falavigna, M. Omologo, "A preliminary statistical evaluation of manual and automatic segmentation discrepancies", in *Proc. of EUROSPEECH-1991*, Genoa: ICSA, 1991, pp. 693–696.

[2] B. M. Wesenick, A. Kipp, "Estimating The Quality Of Phonetic Transcriptions And Segmentations Of Speech Signals", in *Proc. of ICSLP-1996*, Philadelphia:ICSA, 1996, pp. 129–132.

[3] P. Horak, "Automatic Speech Segmentation Based on Alignment with a Text-to-Speech System", *Improvements in Speech Synthesis*, New York:Wiley, pp. 328–338, 2002.

[4] F. Malfrere, O. Deroo, T. Dutoit, C. Ris, "Phonetic alignment: speech synthesis-based vs. viterbi-based", *Speech Communication*, Mons: Elsevier, vol. 4, no. 40, pp. 503–515, 2003.

[5] A. Sethy, S. S. Narayanan, "Refined speech segmentation for concatenative speech synthesis", in *Proc. of ICSLP-2002*, Denver: ICSA, 2002, pp. 149–152.

[6] L. Deng, *Dynamic Speech Models: Theory, Algorithms, and Applications*. Morgan and Claypool, 2006, pp. 4–6.

[7] T. Dutoit, "High-quality text-to-speech synthesis: An overview", *Journal of Electrical & Electronics Engineering*, Institution of Engineers Australia, vol. 1, no. 17, pp. 25–36, 1997.

[8] W. Walker, P. Lamere, P. Kwok, "FreeTTS: a performance case study", Tech Report: TR-2002-114, Sun Microsystems. Inc., 2002, pp. 1–3.

[9] H. Yang, C. Oehlke, C. Meinel, "German Speech Recognition: A Solution for the Analysis and Processing of Lecture Recordings", in *Proc. of ICIS-2011*, Washington:IEEE Computer Society, , 2011, pp. 201–206.

[10] A. Kazlauskiene, K. Velickaite, "Pastabos del lietuviu kalbejimo tempo', *Acta Linguistica Lituanica*, Vilnius: Institute of the Lithuanian Language, no. 48, pp. 49–58, 2003.

[11] H. B. Weil, F. L. Street, "What Is An Adequate Model?", in *Proc. of System Dynamics-1983*, Massachusetts: System Dynamics Society, 1983, pp. 281–321.

[12] J. H. Eggen, "On the quality of synthetic speech evaluation and improvements", Ph.D. Thesis, Technische Universiteit Eindhoven, 1992, pp. 9–15.

[13] A. Anskaitis, "Koduoto balso kokybės tyrimas", Ph.D. Thesis, Vilnius Gediminas Technical University, 2009, pp. 11–19.

[14] D. Gibbon, R. Moore, R. Winski, *Handbook of standards and resources for spoken language systems*, New York :Mouton De Gruyter, 1997, pp. 56–61.

[15] M. Greibus, L. Telksnys, "Rule Based Speech Signal Segmentation", *Journal of Telecommunications and Information Technology*, Warsaw: National Institute of Telecommunications, no. 1, pp. 37–44, 2011.