ZERO-CROSSING-BASED TEMPORAL SEGMENTATION AND CLASSIFICATION OF AUDIO SIGNALS

Konstantin Biatov, Martha Larson, Stefan Eickeler

Fraunhofer Institute for Media Communication (IMK), Sankt Augustin, Germany
{biatov, larson, stefan.eickeler}@imk.fraunhofer.de

ABSTRACT

The task of audio classification is streamlined if the audio stream can be decomposed in real time into homogeneous segments larger than the individual frame. If segment morphology reflects the underlying audio type, classifier design and training becomes more transparent and can directly exploit real-world knowledge. The experiments presented here explore real-time zero-crossing-based audio stream segmentation and classification. We segment an un-edited audio stream from a German-language radio program into three classes: speech, speaker breathing and music. Features related to modulation of the amplitude envelop are exploited to segment the audio stream into homogeneous sub-syllable segments. The segments prove useful for classifying the audio-stream, especially the speech portions. We use features extracted from a zero-crossing-based pseudo-spectrum to train a Gaussian classifier that classifies audio content into the three categories. Our results demonstrate that our zero-crossing-based segmentation and classification method is a viable one, yielding satisfactory performance in real time.

1 INTRODUCTION

Progress of audio processing technologies has led to an expanding role for audio as an information source and a communication medium. The increasing importance of audio data has inspired the vision that that audio should be just as easily segmentable, indexable, searchable and even as translatable as text. Motivated by this vision is the move away from frequency-domain and other calculation intensive approaches, wherever there are not justified by the task or the medium.

This paper is an exploratory investigation in the direction of real-time segmentation and classification of audio data. This work was driven by two central insights. First, audio processing does not have to take place on the frame level, but rather that benefit is to be derived by matching the method to the medium. In our experiments we use modulations in the amplitude envelop to first decompose the audio stream into segments motivated by signal morphology, which in the case of speech turn out to be sub-syllables. Second, audio processing does not have to take place in the frequency domain. We introduce a pseudo-spectrum derived from the distribution of zero-crossings and show that features can be extracted from it that yield reliable classification into audio categories. We segment an un-edited audio stream from a German-language radio program into three classes: speech, speaker breathing and music.

Our experiments draw motivation from the literature on segmentation and classification based on larger-than-frame segments. A two pass classification was introduced in [1]. In the first pass, speech/non-speech boundaries were detected using a reduced phoneme inventory and gender-independent acoustic models. The second pass identified speaker changes, using only the boundaries detected in the first pass as potential candidates. Although this approach increased segmentation accuracy, it also increased segmentation time.

Several approaches have decomposed speech into syllable and syllable-like units. An automatic segmentation of speech into syllable units based on the convex hull of the loudness function in described in [2]. Relative loudness maximums are interpreted as potential syllabic peaks and relative loudness minimums as potential syllabic boundaries. A local loudness minimum separated from another local minimum by less than 100 ms may be insignificant, but a minimum of the same magnitude which had no other minimums within 500 ms would indicate a syllable boundary.

Automatic detection of syllable boundaries based on full-band energy contour and voicing detection using modified autocorrelation is described in [3]. The syllable boundary is determined to be the point of minimum energy contour within each portion of consecutive possibly unvoiced frames.

[4] describes a new representational format for speech, the modulation spectrogram, that represents amplitude modulation frequencies in the speech signal between 0 and 8 Hz, with the peak at 4 Hz corresponding to modulation spectrum of speech. The modulation spectrogram robustly extracts information related with the syllabic segmentation of speech.

The literature on zero-crossing rate is a growing corpus, and also served as an inspiration for our approach. Possibilities for spectral analysis based on zero-crossing are described in [5]. In [6] a spectral analyzer based on zero-crossing is demonstrated. Using zero-crossing interval measurement for formant frequency estimation in noise is presented in [7]. Experiments on speech recognition in a noisy, real–world environment in which the frequency information of the signal is obtained from zero-crossing intervals are described [8]. In [9] experiments which use gravity centers of energy as additional feature to the classical set of MFCC in automatic speech recognition system are described. These experiments demonstrate improved recognition performance when gravity centers are computed from zero-crossing intervals detected at the output of the filters of an ear model.

Using just zero-crossings for speech/music discrimination has been investigated previously in [10]. Other research on speech/music discrimination and classification has used temporal information from speech waveform and as well as spectral information. Spectral information has include such features as 4 Hz modulation energy, rolloff of the spectrum, spectral flux and spectral centroid. In this paper we would like to demonstrate that features like these can be extracted from a zero-crossing-based pseudo-spectrum and used to classify and audio stream with satisfactory accuracy.

In the next section of this paper, section 2, we describe our zero-crossing-based downsampling of the time-signal and present motivation for the approach. We introduce our method for extracting sub-syllable segments by looking at modulations in the amplitude envelop. We detail the calculation of the zero-crossing-based pseudo-spectrum and the features we derive from it. In section 4 we describe our classification experiments. Section 5 presents conclusions and outlook.

2 PSEUDO-SPECTRAL ANALYSIS

2.1 Zero-crossing-based Smoothing

Assuming that signal is normalized, i.e. the mean has been removed, we define zero crossings, t_n, $n = 1,2,3...$, as the times at which the signal changes sign or is equal to zero. The successive zero-crossing intervals are defined as $z_n = t_n - t_{n-1}$. As discussed in [7] the successive zero-crossing intervals of sinusoidal signal exhibit high consistency and are inversely related to frequency. The signal resulting from the sum of more than one sinusoidal component, however, may not satisfy this principle. Usually the speech signal between two zero-crossings has more than one sinusoid. To be sure that the representation of speech signal as one sinusoid between two zero-crossings is valid for the purpose of discrimination we have implemented the method described in [12]. The audio signal is represented as a sequence of successive zero-crossing intervals and the maximal signal amplitudes for those intervals where amplitude is positive and minimal amplitude for the intervals where amplitude is negative. The signal is represented as a sequence (z_n, A_n), $n=1,2,3,....$

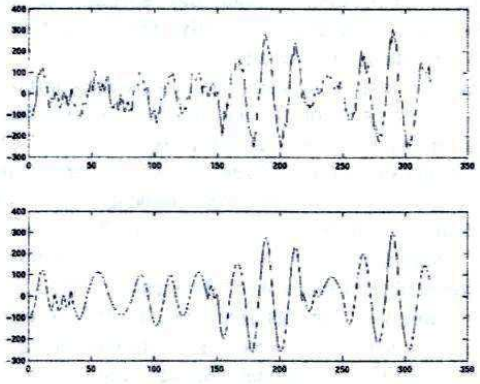

Figure 1. Comparison of original signal and recovered signals

An audio signal restored from this sequence retains sufficient quality to be understood by a human listener, and therefore holds good potential for discriminating speech, music, song, and other non-speech signals. Inconsistencies in the relationship between frequency and zero-crossing rate in the speech signal, remain well contained. From Figure 1 an impression can be gained of exactly how close the original speech signal is to the speech signal restored after zero-crossing-based downsampling.

According to Kedem [5] the zero-crossing rate of repeatedly differentiated series converges to its least upper bound. The restored signal has the same zero-crossing rate as its derivative. Since we are assured of an least upper bound, we are justified in considering that in each successive interval we have only one sinusoid and in using the inverse of the length of the interval as a frequency of the signal in this interval.

2.2 Extracting larger-than-frame segments

The changes that are relevant for segmenting and classifying a radio program, are not those that occur on a frame to frame, but rather those which occur between some higher-level content based units. Audio content processing can be streamlined if a higher-level segment can be found, which is identifiable in the signal, but whose morphology reflects underlying audio content.

Segments which are morphologically justified are also to be homogeneous. We want to exploit facts about the audio stream such as that a strain of music is never going to separate two halves of a syllable of speech. We are interested in morphological properties of the signal that correspond to detectable sections of the audio input. For breathing this might be individual breaths, and for music, the notes. Here we concentrate on extracting a basic unit for spoken audio, namely the syllable. We find that sub-syllables, units that are useful since they never straddle syllable boundaries and can be identified reliably using the minima in the amplitude envelop of the signal.

The signal is divided into frames, each 20 ms seconds long with an overlap of 10 ms. For each frame we extract the maximal non-negative amplitude from the set of amplitudes A_n, $n=1,2,3,....$ corresponding to interval between successive zero-crossing, z_n, A_n $n=1,2,3,....$ On a new signal composed from these maximum amplitudes we find the points which are local minima. These are the boundaries of our larger-than-frame units.

2.3 Deriving the Pseudo-spectrum

The zero-crossing-based pseudo-spectrum is calculated from the restored signal described in sub-section 2.1 using the following procedure. From each 20 ms frame (overlap 10 ms) we extract the lengths of the intervals between successive zero-crossings. Then extracted periods are converted into pseudo-frequencies, by $f = 1/z$, where z is the length of the interval between successive zero-crossings. For each interval between successive zero-crossings we extract the amplitude that is maximal for the interval with positive sign of the signal, and is minimal for the interval with negative sign of signal. Then using 24 frequency bands according to the Bark scale, we group frequencies that correspond to one band together and then normalize all groups using sum of all frequencies. As the result for each 20 ms frame we have smoothed a pseudo-spectrum which can be deployed for a range of discrimination tasks.

The resulting pseudo-spectrum is well smoothed, but retains the discriminative features of the original speech signal. Figure 3 illustrates the pseudo-spectrum corresponding to the 16 kHz waveform depicted in Figure 2. These figures motivate the use

of the pseudo-spectrum distinguish voiced and unvoiced parts of the signal and identify the syllabic modulation of the speech signal. Figure 4 is the pseudo-spectrogram of a waveform in which the speaker breathes between words or phrases. The duration of the breath is 100 ms. Once again, visual inspection confirms that the pseudo-spectrum contains enough information about the signal to allow a discrimination to be made.

Using the amplitudes which correspond to each zero-crossing interval we calculate a value related to energy.

$$S = \frac{abs(z_n * A_n)}{2} \qquad (1)$$

S approximates the square of the amplitude of the envelop corresponding to interval z_n. We group together all S values corresponding to the same Bark band and normalize these values by dividing by total sum for all intervals. In this way we produce a pseudo-spectrum representing the distribution of energy with respect to frequency bands.

We use this pseudo-energy spectrum to calculate an additional feature, the rolloff point. We define our rolloff point to be the frequency below which 85% of the energy in the spectrum is concentrated.

Additionally we use the centers of gravity of the Bark-spectrum frequency bands, proposed in [9].

$$CG = \frac{\sum_{i=1}^{n} A_i z_i}{\sum_{i=1}^{n} A_i} \qquad (2)$$

From the pseudo-spectrum we also calculate the frequency corresponding to the maximum zero-crossing interval in each frame. The final feature that we derive from the pseudo-spectrum is the Euclidean distance between the pseudo-spectral vectors of neighboring frames.

3 EXPERIMENTS

We experimented on data from our *Kalenderblatt* database, which contains radio programs recorded from the Deutsche Welle *Kalenderblatt* series and the accompanying transcripts. The programs are five minutes in length and each contain about 650 running words. They are liberally interspersed with music and other sound effects. Although each program is narrated by a single speaker, many other voices are present in the form of interviews and original sound footage. The radio programs were downloaded from the Internet (http://www.kalenderblatt.de) and resampled to 16 kHz. The corresponding text transcripts were also downloaded and normalized. The database is described in detail in [14]. Also in [14] is a description of the semi-automatic alignment visualization tool with which we create reference segmentations and segment labels for our experiments.

For the exploratory experiments presented here, we indiscriminately chose one 5-minute program for training and one for testing. We segmented and labeled these with the semi-automatic method down to the syllable level, also marking breathing and music. In this way we generated high-quality reference labels for training and testing. The duration of training data for the experiments presented here is 5 minutes; the duration of test data is also 5 minutes.

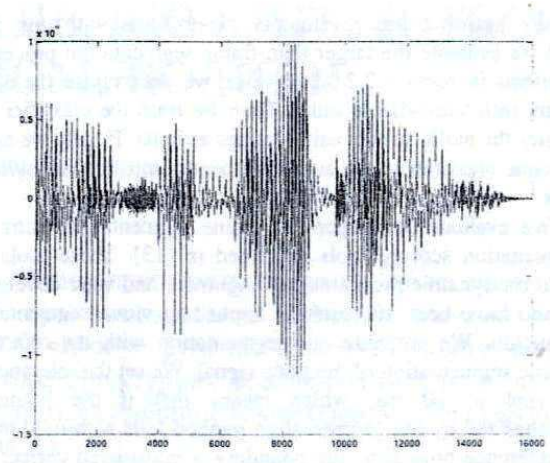

Figure 2. Waveform of speech signal. (20 ms frame with 10 ms step represented on x-axis)

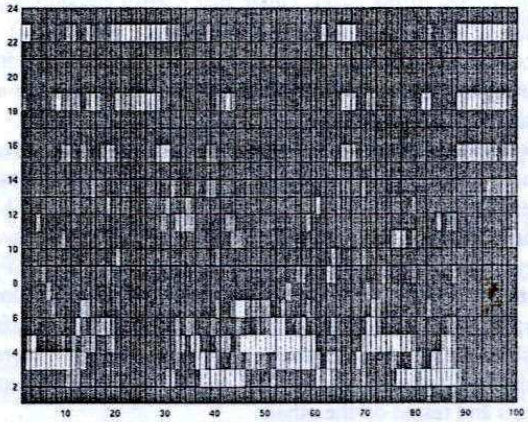

Figure 3: Pseudo spectrum for the waveform depicted in Figure 2. (Bark-scale pseudo-frequencies vs. frames)

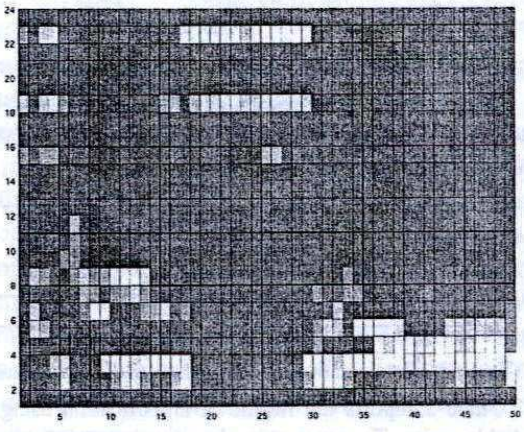

Figure 4: Pseudo spectrum for the waveform containing speaker breathing (Bark-scale pseudo-frequencies vs. frames)

We perform three evaluatory experiments with this data. First we evaluate the larger-than-frame segmentation procedure described in section 2.2 with which we decompose the audio stream into sub-syllable units. Then we train the classifier and classify the audio stream using frames as units. Finally we apply the same classifier to the audio segmented into the sub-syllable units.

We evaluate our larger-than-frame segmentation using the segmentation scoring tools described in [13]. These tools are based on dynamic programming alignment and were developed for and have been successfully applied to video segmentation evaluation. We compare our segmentation with the reference syllable segmentation of the same signal. We set the tolerance of the tool to 20 ms, which means that if the boundary hypothesized by our segmentation method falls within 20 ms of the reference boundary, the boundary is considered correct. 20 ms represents a conservatively narrow margin of tolerance, since an average syllable is 250 ms in length. Our segmentation scores 87.5% insertions and 12.5% deletions with respect to the reference segments. This means that our segments in most cases are proper sub-units of syllables and cross syllable boundaries in only 12.5% of the cases. Given that correct syllabification in German is partially a matter of convention, we feel that these results indicate that our larger-than-frame segments are indeed capturing a sub-unit that provides a useful link between signal structure and audio content.

For classification we use a maximum likelihood multivariate Gaussian classifier. The input vectors for the classifier consist of the features extracted from the pseudo-spectrum (described in section 2.3) plus their variances. These features are: average energy, gravity centers of pseudo-frequency bands, rolloff point of energy distribution, frequency corresponding to the longest interval between successive zero-crossings present in frame, Euclidean distance between pseudo-spectral vectors of neighboring frames. The classifier was trained on one 5 minute program and tested on the other.

	classified as music	classified as breath.	classified as speech
music data	**84.2%**	3.7%	11.1%
breathing data	8.2%	**89.2%**	2.6%
speech data	25.6%	12.2%	**62.2%**

Table 1: Classification on frames (20 ms, with 10 ms step)

Table 1 reports the results of classification using frames as the units. Music and breathing are classified better than speech using frame as the basic classification unit.

	classified as music	classified as speech
music data	**68%**	32%
speech data	6%	**94%**

Table 2: Classification on sub-syllables (larger-than-frame)

Table 2 reports the results of classification using the sub-syllable units. To score this classification, we look at which frames are contained in the sub-syllable and see which class label is associated with the majority of them in the reference segmentation. Speech is classified better than music using the sub-syllable units. We feel that this result reflects the utility of choosing a larger-than-frame unit that is related to the morphology of the audio content.

4 CONCLUSIONS

The zero-crossing pseudo-spectrum proposed here has proven itself to be a valuable source of features for real-time audio classification and has been shown to allow effective discrimination between speech, speaker breathing and music. Further results suggest that larger-than-frame units enhance classifier performance, but that they must be well-matched with the audio content.

Future work will focus on identifying additional features that can be derived from the pseudo-spectrum that might aid classification of audio, particularly into classes above and beyond the three we experiment with here. Finally, we hope that further work will allow us to develop a larger-than-frame segment that will be as useful for music and breath classification as the sub-syllable units are for speech classification.

5 REFERENCES

[1] Makhoul, J. et al. (2000). Speech and Language Technologies for Audio Indexing and Retrieval: Advances and Applications. Proceedings of the IEEE, 88:1338-1353.

[2] Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. Journal of the Acoustic Society of America, vol. 58, No. 4, pp.880-883.

[3] Sakaguchi, S., Arai, T. and Murahara, Y. (2000). The effect of polarity of speech on human perception and data hiding as an application. ICASSP 2000.

[4] Greenberg, S. and Kingsbury, B. (1997). The modulation spectrogram: in pursuit of an invariant representation of speech. ICASSP 1997, pp. 1647- 1650.

[5] Kedem, B. (1986). Spectral analysis and discrimination by zero-crossing. Proceedings of IEEE, VOL. 74, No. 11, pp. 1477-1493.

[6] Kay, S. A zero-crossing based spectrum analyzer. (1986). IEEE Transaction on Acoustic, Speech and Signal Processing, vol. ASSP-34, No. 1, pp. 96-104.

[7] Sreenivas, T. and Niederjohn, R. (1992). Zero-crossing based spectral analysis and SVD spectral analysis for formant frequency estimation in noise. IEEE Transaction on Signal Processing, vol. 40, No. 2, pp. 282-293.

[8] Kim, D.S., Lee, S.Y. and Kil, R.M. (1999). Auditory processing of speech signal for robust speech recognition in real-world noise environment. IEEE Transaction on Speech and Audio Processing, 7, No. 1, pp. 59-69.

[9] De Mori, R., Moisa, L., Gemello, R., Mana, F. and Albesano, D. (2001). Augmenting standard speech recognition features with energy gravity centres. Computer Speech and Language, 15, pp. 341-354.

[10] Saunders, J. (1996). Real-Time Discrimination of Broadcast Speech/Music", Proc. ICASSP 1996, pp. 993-996.

[11] Scheirer, E. and Slaney, M. Construction and evaluation of a robust multifeature speech/music discriminator. ICASSP 97, vol. 2, pp. 1331-1334.

[12] Vintsiuk, T., Kulias, A. and Dys, A. (1982). A. economic presentation of acoustic signal in computer. ARSO-12, Institute of Cybernetics of AS USSR.

[13] Eickeler, S. and Rigoll, G. (2000). A novel error measure for the evaluation of video indexing systems. ICASSP 2000.

[14] Eickeler, S., Larson, M., Rüter, W. and Köhler, J., (2002) Creation of an Annotated German Broadcast Speech Database for Spoken Document Retrieval," to appear, Proceedings of LREC2002.

ДЕТЕКТОР ЗВУКОВИХ ОБРАЗІВ, ЩО ЕМІТУЮТЬСЯ СЛУХОВИМ ОРГАНОМ ЛЮДИНИ

В.Г. Грибенко, *інженер звукотехніки, м. Яготин Київської обл.*
abris 2001 @ hotmail. com.

Анотація.

Явище емісії звукових коливань периферійним слуховим органом людини залишається загадковим у нейрофізіологічному сенсі [1]. Дослідження Кемпа [2, 3] стимульованої та спонтанної отоакустичної емісії започаткували вивчення механізму цього явища.

Для кращого розуміння походження сигналів, генерованих слуховими центрами мозку людини в процесі мислення, та їх перетворення і трансляції, автор пропонує новий підхід у дослідженні механізму слуху людини на порозі чутності [4] та бачення слухового органу не тільки як пасивного приймача зовнішніх слухових коливань, але і активного органу емісії звукових образів, що відповідають думкам людини.

Пропонується ідея іноваційного проекту створення оптоелектронного отоендоскопа – приладу, за допомогою якого стане можливою селекція інформації з шуму коливань барабанної перетинки та ідентифікація її з мовними музичними та іншими звуковими образами, що відповідають певній розумовій діяльності.

Вступ.
Нейрофізіологічна версія розпізнавальної функції слухових структур мозку людини

В анатомії мозку тварин розрізняють декілька слухових нервових шляхів, що місцями перетинаються між собою [5].

Висхідні шляхи – це послідовна сукупність асоційованих нервових слухових утворень з тонотопічною організацією від органа Корті до височних ділянок кори.

Ці шляхи мають численні зв'язки з ретикулярною формацією мозку – сукупністю структур в центральних його відділах, які регулюють рівень збудження відповідних ділянок центральної нервової системи з корою великих півкуль включно.

Зворотні шляхи (на жаль маловивчені) починаються з відповідних ділянок кори і через нижчі відділи слухової системи з'єднуються з органом Корті [6].

Аналізуючи слухову систему мозку (ССМ) людини можна припустити наявність селективного зворотного зв'язку між органом Корті і відповідними відділами мозку.

Автор припускається думки, що на порозі чутності при розпізнаванні зовнішніх звукових образів ССМ виконує функцію багатоканального надрегенератора структурних фрагментів імпульсів, що відповідають внутрішнім усвідомленим слуховим образам (рис.1).

Ці фрагменти імпульсів після селекції та адресації в ССМ попадають в орган Корті, де перетворюються у відповідний спектр механічних коливань, які через внутрішній і середній відділи вуха передаються на поверхню барабанної перетинки (ПБП) у вигляді фрагментів внутрішніх, але вже звукових образів.

За відсутності зовнішніх звукових образів ПБП під дією вушного шуму постійно знаходиться у коливальному стані. Спектральні характеристики цього шуму контролюються ССМ [7].

ССМ, маніпулюючи довжиною фрагментів слухових імпульсів, їх інтенсивністю, формою, місцем та часом генерації через функцію органа Корті підлаштовують рух ПБП синхронно її коливанню, збудженому дією фрагмента зовнішнього звукового образу на порозі чутності, чим збільшують його амплітуду, полегшуючи розпізнавання.

При асинхронній модуляції руху ПБП фрагментами зовнішнього і внутрішнього звукових образів, інформація про невідповідність динаміки коливань ПБП динаміці фрагмента внутрішнього звукового, а значить і слухового образу, перетворюється і передається до ССМ органом Корті, чим викликається негайна корекція слухового фрагмента, що сприяє розпізнаванню та усвідомленню образа.

Насамкінець, можливо припустити, що в стані роздумів у "повній тиші" (Nзовн.шуму ≤ 0 дБ) на ПБП з боку ССМ надходять лише звукові коливання, обумовлені думкою та іншою життєдіяльною та патологічною емісією (т.ч. ми "слухаємо" свої думки).

Вищенаведені припущення дещо полегшують розуміння феноменально високої локальної чутливості вуха в діапазоні частот 1 – 5 кГц, бо чутливість до синхронного збурення динамічних систем суттєво вища, ніж статичних.

Принцип роботи детектора власних коливань барабанної перетинки та вимоги до його конструкції

Подумки поставимо себе перед необхідністю визначити у темному місці дотиком динаміку рухомого предмета і його вібрацію.

При випадковому короткочасному дотику результат напевне буде невизначеним. Задача спрощується, якщо рукою супроводжувати предмет, не порушуючи траєкторії і режиму його руху.

Подібну механічну аналогію можливо розповсюдити на принцип розпізнавання структури звукових коливань мембрани за допомогою рухомого на відповідній відстані від ПБП детектора.

Необхідність гранично високої чутливості детектора до амплітуди вібрації ПБП (атомного масштабу) і водночас достатності його динамічного діапазону для компенсації значних амплітуд неінформативних звукових збуджень від процесів дихання, серцево-судинної діяльності, стоматофонової та шлункової активності, а також інших спонтанних звукових реакцій ПБП обумовлена слідуючими головними засадами до проекту детектора:

перетворювання фізичних параметрів у датчику повинно бути механо-оптичним задля уникнення значного біоелектронного шуму в зоні розташування перетворюючого елемента детектора;

торець мікросвітловода необхідно покрити напівпрозорим дзеркальним шаром металу, мікрооб'єм тіла барабанної перетинки безпосередньо в зоні торця та проміжок між ними при вібрації ПБП повинні створювати модулююче промінь світла середовище, тобто мікрооб'єм тіла ПБП в її центрі повинен бути елементом конструкції перетворювача в сенсі оптимальності його використання як модулюючого відбите світло середовища.

Фізичний процес в оптичному мікроконтакті грунтується на природній модуляції енергії індукованого випромінювання коливанням ПБП у квазірезонаторі, створеному напівпрозорим дзеркалом торця мікросвітловода та відбиваючим частину світлового променя об'ємом на ПБП.

Коректне апаратурне втілення такого принципу зробить можливим створення мікрооптичного контакту з необхідним ступенем модуляції відбитого від ПБП світла.

Оптимальний режим роботи мікрооптичного контакту визначиться пороговим рівнем автоматичного утримання його параметрів, а також значенням інтенсивності опорного променя без ризику деструкції мікроконтактної зони на ПБП.

Слід зазначити, що запропонований до розгляду детектор повинен бути невід'ємною частиною мультипроцесорної системи оброблення та розпізнавання динамічних образів з граничною якістю специфікацій.

Ескіз пілот-конструкції детектора отоендоскопа

Пара детекторів власних коливань ПБП складає вхідну частину отоендоскопа і має бути симетрично розміщена на звукопоглинаючому оголів'ї згідно попередньо одержаних топограм зовнішніх слухових каналів (ЗСК) голови пацієнта.

Наступний процес юстирування полягає у приведенні детекторів в положення (див. рис.2) відносно зовнішньої поверхні барабанної перетинки 1, при якому між ними виникає мікрооптичний контакт.

В ЗСК входить гнучкий звукопоглинаючий штуцер 2 з трьома капілярами 2.1 і пневмопорожнинами 2.2 на їх кінцях, утвореними вклеєним в штуцер 2 платиноіридієвим розрізним з антифрикційною обробкою внутрішньої його поверхні циліндром 2.3 та неприклеєними до нього ділянками матеріалу штуцера 2 в кінці кожного капіляра 2.1.

Таким чином, на кінці штуцера 2 утворюються пневмопорожнини, які окремо регулюються по висоті, що дає можливість фіксувати штуцер 2 у необхідному положенні згідно параметрів ЗСК.

Всередині штуцера 2 вздовж його осі знаходиться рухомий зонд, який складається з гнучкого електропровідного і звукопоглинаючого штока 3, з'єднаного з платиноіридієвим наконечником 4; п'єзоелемента 5, закріпленого електропровідним цементом в фасонному пазі наконечника 4; та оптоволоконного трійника 6 світловода.

Шток 3 і наконечник 4 мають співпадаючі канавки, в яких залягає трійник 6 світловода [8], а також ізольований провідник 5.1 для подачі напруги на п'єзоелемент 5.

Закріплена на п'єзоелементі 5 частина 6.1 трійника має конічну форму і радіальний градієнт показника заломлення, а потім переходить у короткий відрізок одномодового волоконного світловоду, торець якого покритий напівпрозорим дзеркальним шаром металу. Мікрооптичний контакт створюється між торцем частини 6.1 трійника і ПБП.

Частина 6.2 трійника підключена до світловипромінюючого діода 7, закріпленого в канавці штока 3, а інформаційна частина 6.3 трійника – до мікроканального ФЕП – дінода 8 (далі по тексту дінода 8).

Шток 3 і світловипромінюючий діод 7 приклеєні до голчастого радіатора 9, який рухається трьома реверсивними лінійними п'єзоелектричними мікродвигунами 10.

Таким чином, зонд зміщується з робочої точки в неробочу і навпаки в штуцері 2 як в пеналі – запобіжнику від пошкодження частини 6.1 трійника.

Радіатор 9 виконаний як тепловідвід термоелектричного мінікріоблока 11, який охолоджує дінод 8, вмонтований в колектор 11.3 холодних спаїв.

На радіаторі 9 закріплена металічна плата 12 з двобічним монтажем екранованих мікропотужних НВІС – процесорів, обслуговуючих датчик.

Мікродвигуни 10 вклеєні в корпус-кондуктор 13 з трьома карпілярами 13.1 сумісними з капілярами 2.1.

Дінод 8 має в центральній своїй частині 8.1 металокерамічний спай 8.2 для її температурного обмеження.

Місце припайки керамічних анодної 8.3 та катодної 8.4 частин дінода 8 до центральної його

Рис. 1. Структурно-функціональна схема системи виявлення звукових сигналів, що емітуються органом Корті

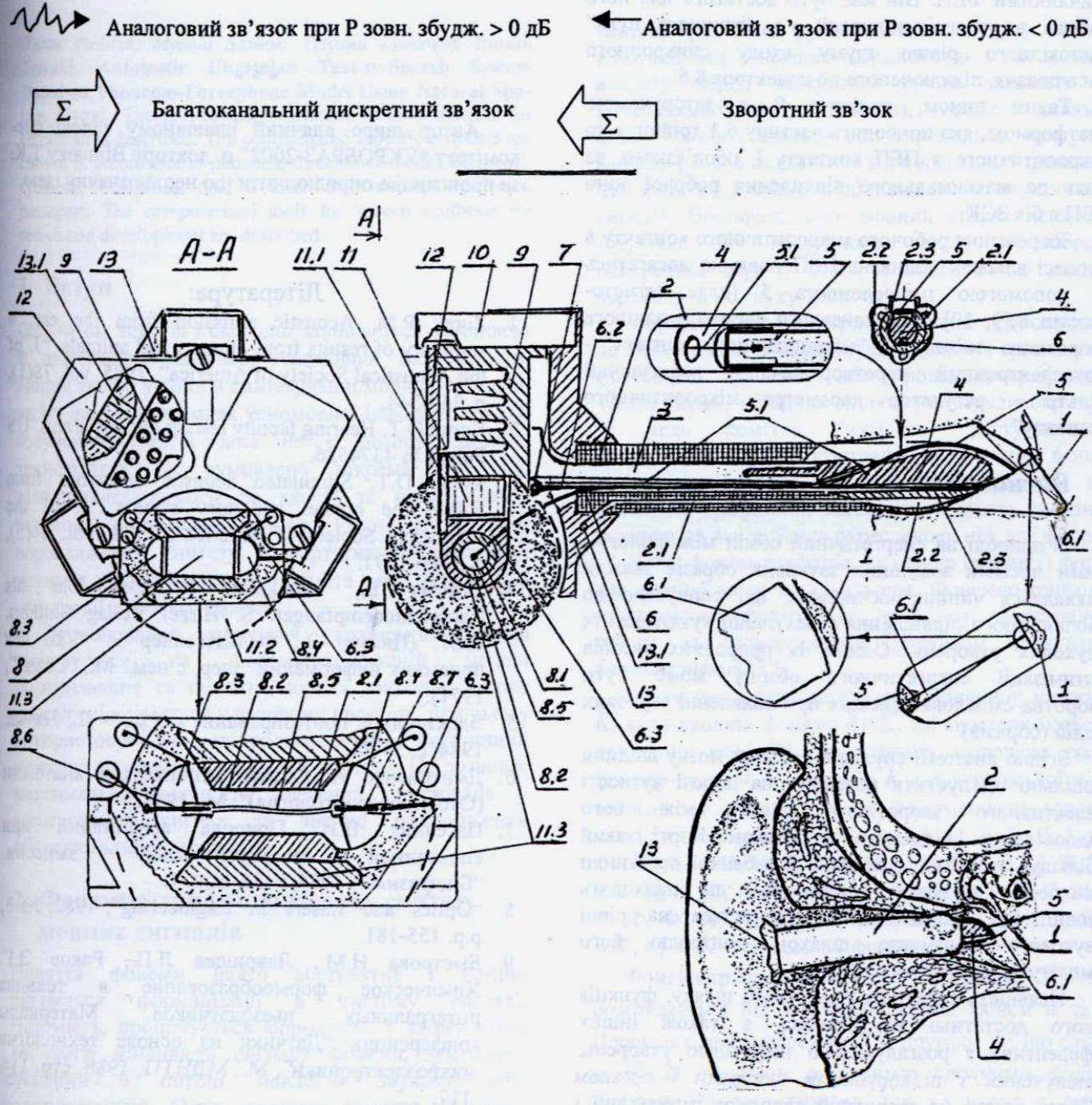

Рис. 2. Конструкція датчика отоендоскопа.

1 – барабанна перетинка; 2 – штуцер (2.1 – капіляр; 2.2 - пневмопорожнина; 2.3 – розрізний циліндр); 3 - шток; 4 - наконечник; 5 – п'єзоелемент (5.1 – провідник); 6 – трійник світловода (6.1 – приймальна частина; 6.2 – опорна частина; 6.3 – інформаційна частина); 7 – світловипромінюючий діод; 8 – дінод (8.1 – центральна частина; 8.2 – металокерамічний спай; 8.3 – анодна частина; 8.4 – катодна частина; 8.5 – мікроканал; 8.6 – колектор електронів); 9 – радіатор; 10 – п'єзоелектричний двигун; 11 – мінікріоблок (11.1 – комутаційна панелька; 11.2 – зовнішній теплоізолятор).

частини 8.1 є відповідно анодом і катодом, які електрично з'єднані між собою покриттям мікроканала 8.5.

Колектор електронів 8.6 електроізольований від анода 8.3.

В катодну частину 8.4 впаяний світловод 8.7.

Коефіцієнт посилення фотоструму дінода 8 не слід робити занадто високим в порівнянні із звичайними ФЕП. Він має бути достатнім для його роботи в режимі модуляції т.з. "жирного нуля" відповідного рівню шуму входу синхронного посилювача, підключеного до колектора 8.6.

Таким чином, радіатор 9 є авторухомою платформою, яка приводить частину 6.1 трійника до мікрооптичного з ПБП контакту і, обов'язково, за мить до максимального відхилення робочої зони ПБП в бік ЗСК.

Збереження робочого мікрооптичного контакту в процесі власних коливань ПБП повинно досягатись за допомогою п'єзоелемента 5 (клас сегнето-еластиків)[9, 10] як виконавчого елемента ланцюга зворотного зв'язку – "мікрооптичний контакт – фотоелектронний перетворювач – коригуючий фільтр – регулятор параметра мікрооптичного контакту".

Висновки.

В природі на енергетичний обмін між біологічними носіями зовнішніх звукових образів завжди накладався чинник виживання, що спонукало до еволюційного підвищення селективної чутливості їх слухових утворень. Одним із природних засобів оптимізації енергетичного обміну може бути зворотна смислова селекція при виявленні звукових подій (образів).

Згідно анатомії слухової системи мозку людини доцільно припустити наявність на порозі чутності селективного зворотного зв'язку між його відповідними відділами і органом Корті, який збільшує амплітуду коливань барабанної перетинки синхронно звуковим коливанням, що надходять ззовні. Це збільшення забезпечується на рівні звукового фрагмента шляхом контролю його амплітуди та швидкості її зміни.

Наявність складного зворотного шляху, функція якого достатньо не вивчена, а також інших аферентних з розгалуженою інервацією утворень, сполучених з підкорковими центрами і органом Корті, вказує на зворотний напрямок імпульсації і, не виключено, що вона пов'язана також із свідомою вербальною діяльністю мозку. Відповідь на це питання можуть дати тільки точні експерименти.

Виявлення спонтанної отоакустичної емісії стало можливим завдяки впровадженню електретних мікрофонів, але принципова наявність значних шумів стримує їх використання в сенсі підвищення чутливості відповідного експериментального обладнання. Тому був запропонований більш чутливий метод виявлення власних коливань зовнішньої поверхні барабанної перетинки, започаткований на способі модуляції нею інтенсивності світлового променя, тим більш, що сучасний рівень значної низки технологій дозволяє провести репрезентативне дослідження тонкого спектру власних коливань барабанної перетинки в діапазоні збуджень на порозі чутності.

* * *

Автор щиро вдячний шановному голові орг-комітету "УКРОБРАЗ-2002" п. доктору Вінцюку Т.К. за пропозицію оприлюднити цю неоднозначну ідею.

Література:

1. Zurek P.M. Acoustic emission from the ear: a summary of results from humans and animals. "J. of the Acoustical Society of America", 1985, vol. 78(1), p. 340-344.
2. Kemp D.T. Hearing faculty testing and apparatus. US Patent № 4374526.
3. Kemp D.T. Stimulated acoustic emissions from within the human audiotory system "J. of the Acoustical Society of America", 1978, vol. 64(5), p. 1386-1391.
4. Zwicker E., Feldtkeller R. Das Ohr als Nachrichtenempfanger. S. Hirzel Verlag Stuthart. 1967. (Цвикер Э., Фельдткеллер Р. "Ухо как приемник информации", пер. с нем.. М., "Связъ", 1971).
5. Электродное протезирование слуха. – Л., Наука, 1984, с. 53-80.
6. Дмитриенко И. "Атлас клинической анатомии (Оториноларингология)", М., 1998.
7. Пасечник В.И. Природа флуктуаций при спонтанной отоакустической эмиссии. "Биофизика", 1984, т. 34, вып. 1.
8. "Optics and Lasers in Engineering", 1982, № 3, p.p. 155-181.
9. Быстрова Н.М., Лаврищев Л.П., Раков Э.Г. Химическое формообразование в технике интегральных пьезодатчиков. Материалы конференции "Датчики на основе технологии микроэлектроники", М., МДНТП, 1986, стр. 114-117.
10. J. van Randaraat, Setterlington R. E. Piezoelectric Geramics. Eds. Eindhoven: N.V. Philips, 1974.

АВТОМАТИЧНИЙ ОЗВУЧУВАЧ УКРАЇНСЬКИХ ТЕКСТІВ НА ОСНОВІ ФОНЕМНО-ТРИФОННОЇ МОДЕЛІ З ВИКОРИСТАННЯМ ПРИРОДНОГО МОВНОГО СИГНАЛУ

Тарас Вінцюк, Микола Сажок, Тетяна Людовик, Руслан Селюх

Міжнародний науково-навчальний центр інформаційних технологій та систем
40 просп. Академіка Глушкова, Київ 03680
Електронна пошта: {vintsiuk, mykola, tetyana_lyudovyk, selyukh}@uasoiro.org.ua

Taras Vintsiuk, Mykola Sazhok, Tetyana Lyudovyk, Ruslan Selyukh. **Automatic Ukrainian Text-to-Speech System Based on Phoneme-Threephone Model Using Natural Spoken Signal.** The text-to-speech system in time domain for Ukrainian is described. The concatenated acoustic elements are chosen in accordance to phoneme-threephone model for speech synthesis. Acoustical data is taken from the speaker voice passeport. The computerized tools for speech synthesis research and development are described.

1 Вступ

Озвучення текстів і сьогодні залишається важливою й актуальною задачею усномовної інформатики. Попри значні успіхи в синтезуванні мовних сигналів, як і в інших напрямках усномовної інформатики [1], озвучення текстів досі не є розповсюдженою технологією. Це зумовлено зокрема тим, що підвищились вимоги до якості та натуральності звучання синтезованої мови. Разом з тим для впровадження синтезу в портативних пристроях повинна задовольнятися вимога на обмеження швидкодії та обсяги пам'яті.

Пропонований озвучувач українських текстів засновується на фонемно-трифонній моделі розпізнавання та синтезу мовних сигналів, синтез сигналу відбувається у часовому просторі, при цьому використовується природномовний акустичний матеріал з усномовного файлу диктора. Це дозволяє максимально зменшити внесення спотворень у згенерований сигнал та значно розвантажити обчислювальний модуль.

2 Фонемно-трифонна модель синтезу мовних сигналів

Поняття фонеми надто абстрактне і погано піддається формалізації в чистому вигляді. Натомість, пропонується формалізм, у якому взято до уваги мінливість сигналу фонеми, зумовлену сусідніми в потоці мовлення звуками або коартикуляцією. Отже, вводимо поняття фонеми-трифона, коли розглядається фонема в контексті з попередньою та наступною фонемами. Поняття фонеми-трифона покладається в основу акустичних моделей як розпізнавання, так і синтезу усної мови.

Фонемно-трифонна акустична модель усної мови дозволяє врахувати явище коартикуляції, що виникає при взаємодії звуків у потоці мовлення. Справді, при відтворюванні послідовності звуків, що відповідають певним фонемам, рухи мовного апарату людини відбуваються з певною інерційністю. Стан мовного апарату перед наступним рухом залежить від попереднього звуку (фонеми), а отже і динаміка рухів мовного апарату при переході до наступної фонеми різна в залежності від попередньої фонеми, а це, зрештою, якісно відбивається на акустичному сигналі. Водночас, весь мовний апарат немов би готується до наступного звуку, і завершує попередній звук у стані, з якого більш вигідно переходити до звуку, який слідує.

Формалізм фонеми-трифона, який пропонується, дозволяє також конкретизувати поняття границь між фонемами, точніше, між фонемами-трифонами. Таким чином, незначне відлуння попереднього звуку та ледь помітна "чутність" наступного є допустимими при розставлянні границь фонем-трифонів. Уцілому, розставляння границь між фонемами-трифонами залишається досить складною задачею, та все ж більш окресленою, ніж для фонем.

Набір фонем-трифонів, як і набір (алфавіт) фонем для кожної мови є унікальним. Фонемно-трифонна транскрипція формується на основі фонетичного тексту за універсальним правилом стикування фонем-трифонів [2].

Виходимо з того, що задано скінченну множину K, куди входять фонеми $k \in K$, які спостерігаються в природній мові [3]. До алфавіту включено також фонему-паузу #. У множині K для української мови розрізняємо наголошені та ненаголошені голосні, м'які та тверді приголосні: $k \in$ {А, О, У, Е, И, І, А1, О1, У1, Е1, И1, І1, Б, Б', В, В', Г, Г', Ґ, Ґ', Д, Д', Ж, Ж', З, З', Й, К, К', Л, Л', М, М', Н, Н', П, П', Р, Р', С, С', Т, Т', Ф, Ф', Х, Х', Ц, Ц', Ч, Ч', Ш, Ш', ДЗ, ДЗ', ДЖ, ДЖ', #} $\equiv K$ – загалом 57 фонем.

Фонема-трифон $t=uWv$ є фонемою W, яка розглядається під впливом сусідніх фонем u та v. Першою є u, що передує W, а другою – v, що слідує за W. За правилом допустимих сполучень фонем-трифонів [2] допустимими для з'єднання є лише фонеми-трифони вигляду $t_1=uWv$ і $t_2=wVz$ через Wv та wV.

Загальна кількість фонем-трифонів у алфавіті теоретично дорівнює кількості базових фонем у степені три (125000 для алфавіту з 50 фонем). Практично ж можна обмежитися декількома тисячами фонем-трифонів. Відсутні фонеми-трифони

замінюються найближчою згідно з фонемно-трифонною ієрархією.

З природи усномовного сигналу випливає, що дзвінкі фонеми можна описувати (транскрибувати) як послідовність одно-квазіперіодичних мікросегментів, що мають певну форму звукової хвилі та довжину (Рис. 1). Такий опис поширюємо також і на глухі (шумні) фонеми і називатимемо акустичною транскрипцією фонеми-трифона. Послідовність хвиль одноквазіперіодичних мікросегментів утворює акустичний образ (прототип) фонеми-трифона. Сукупність всіх акустичних образів усіх фонем-трифонів, властивих певній людині, складає усномовний файл диктора [2].

Розглянемо процес автоматичного озвучення тексту. Спочатку, проводиться розбиття тексту за схемою: абзац—речення—синтагма (інтонаційна група)—ритмогрупа (група наголосу)—фонетичне слово. З абзаців виділяються речення, кожне речення розбивається на синтагми, ті, в свою чергу, – на ритмогрупи. І вже всередині ритмогрупи визначаються фонетичні слова, які або збігаються з орфографічними словами або містять їх декілька, включно зі службовими словами. Потім з використанням фонетичних знань про усну мову орфографічний текст перетворюється на фонетичну та одночасно й на фонемно-трифонну транскрипції за універсальним правилом. Далі згідно розпізнаних типів синтагм будується інтонаційний контур, розраховуються тривалості поточних одноквазіперіодичних мікросегментів та фонем уцілому. При цьому враховується, що кожна синтагма містить лише один основний наголос, який відповідає ядерній ритмогрупі. Решта ритмогруп – початкова, перед'ядерні та післядерні.

На підставі розрахунків приступаємо до власне формування (синтезу) усномовного сигналу. Для кожної фонеми-трифона з фонемно-трифонної транскрипції обирається акустичний образ (прототип) з бази даних. Цей прототип піддається перетворенням згідно розрахованої тривалості фонеми-трифона та інтонаційного контуру. В результаті цих перетворень маємо отримати визначену перед цим кількість квазіперіодів розрахованої довжини. Перетворені прототипи квазіперіодів і загалом фонем-трифонів об'єднуються, і отриманий в результаті сигнал подається на засоби озвучення.

Визначальною рисою кожної технології синтезу усномовного сигналу є алгоритм змінювання просодичних характеристик прототипів елементів компіляції. Основною метою при зміні просодики, тобто інтонації та темпу, є досягнення якомога кращої якості синтезу за найменших обчислювальних витрат на кожну дискрету синтезованого сигналу.

Як видно з аналізу відповідних алгоритмів [4], у часово-амплітудній області досягається і те й інше. Так, вельми прийнятна якість синтезу за технологією *PSOLA* досягається за досить скромних витрат на обчислення при зміні інтонаційних характеристик прототипу елемента компіляції – до 9 арифметичних дій на одну дискрету. У технології *MBROLA* обчислювальні витрати ще скромніші – 6 арифметичних дій на одну дискрету, і це при кращих показниках якості синтезу, ніж у *PSOLA*. Технології *Unit-Selection* взагалі не передбачають яких-небудь інтонаційних змін прототипу сигналу, хоч при цьому і виникають надмірні витрати пам'яті на зберігання всіх можливих інтонаційних проявів кожного елементу компіляції, що конкатенуються без жодних перетворень сигналу. Це в свою чергу позбавляє певної гнучкості саму систему синтезу усної мови.

Пропонується ще один спосіб компіляції одно-квазіперіодичних мікросегментів у амплітудно-часовій області. В основі цього методу закладено модель лінійного прогнозування сигналу, що дозволяє за певною кількістю попередніх відкліків сигналу спрогнозувати наступні з достатньо високою точністю апроксимації:

$$\tilde{f}_n = -\sum_{s=1}^{s=m} a_s f_{n-s} + \varepsilon_n , \qquad (1)$$

де $\tilde{f}_n$ — відліки прогнозованого сигналу, f_n — відліки спостережуваного сигналу, a_s, $s = 1{:}m$ — параметри передбачення, кількість яких m обирається в межах від 10 до 20, ε_n — похибка прогнозування. Параметри передбачення оцінюються на інтервалі аналізу рівному одному або двом квазіперіодам шляхом, наприклад, мінімізації суми квадратів похибки прогнозу.

Отже, нехай розрахуок тривалостей фонем та довжин квазіперіодів або мікросегментів і їх кількості в кожній фонемі вже виконано. Тоді відповідний прототип фонеми-трифону необхідно піддати темпоральним змінам, тобто привести до розрахованої довжини. Це досягається шляхом доведення кількості квазіперіодів до розрахованої. Таким чином, темпоральні зміни прототипу фонеми-трифону здійснюються шляхом викидання або повторення певних мікросегментів. У свою чергу, інтонаційні зміни сигналу здійснюються внаслідок скорочення або збільшення довжин відповідних квазіперіодів.

При темпоральних змінах, які вимагають подовження або скорочення тривалості прототипу фонеми-трифона до заданої довжини, обчислюється кількість мікрофонем (квазіперіодів), на яку їх необхідно збільшити або зменшити. Збільшення кількості мікрофонем відбувається шляхом повторення деяких квазіперіодів прототипу фонеми-трифону певну кількість разів. Зменшення кількості квазіперіодів здійснюється за рахунок викидання певних квазіперіодів.

Визначення квазіперіодів (мікрофонем), які повторюємо або викидаємо при темпоральних змінах, не є однозначним і потребує додаткових досліджень. Тому маємо право обрати одне з

простіших рішень, а саме повторюємо (викидаємо) переважно центральні квазіперіоди.

Як зазначалося, інтонація сигналу регулюється шляхом задання певної довжини квазіперіодів на вокалізованих ділянках синтезованого сигналу. Для того, щоб подовжити окремо взятий квазіперіод згідно моделі лінійного прогнозування, достатньо знати певну кількість попередніх відліків та значення коефіцієнтів прогнозування на відповідному інтервалі аналізу. Решту відліків, що доповнюють квазіперіод до потрібної довжини, обчислюємо за алгоритмом (1). Скорочення довжини квазіперіоду до заданої здійснюється шляхом видкидання відліків квазіперіоду понад задану довжину.

3 Комп'ютерні засоби дослідження та розроблення фонемно-трифонної моделі синтезу мовлення

Розроблено комп'ютеризовані засоби формування бази даних і знань, які використовуються як для синтезу, так і для пофонемного розпізнавання усної мови. З їх допомогою також проводяться експериментальні дослідження розпізнавання та синтезу.

При формуванні баз даних і знань, опрацюванні методів та алгоритмів, що стосуються усномовної інформатики, розроблені програмні засоби забезпечують виконання таких дій:

- введення фонетичних специфікацій, що включають опис базових фонем та перелік можливих інтонацій для обраної природної мови;
- накопичення навчальної вибірки за заданим текстом;
- автоматична сегментація навчальної вибірки на одноквазіперіодичні мікросегменти та на квазіперіодичні й неперіодичні сегменти;
- автоматизована сегментація навчальної вибірки на фонеми-трифони;
- дослідження та розроблення методів та алгоритмів синтезу усної мови.

Початкова інформація про природну мову складається з алфавіту базових фонем та переліку інтонацій. Задаємо її в режимі діалогу. Спочатку вводимо потрібну мову до списку підтримуваних *мов або активізуємо відповідну мову, якщо вона вже* є в реєстрі. В окремій секції задається алфавіт базових фонем та опис кожної з них. Передбачено підтримку паралельних алфавітів у кодуванні як кирилицею, так і латинськими літерами. Окреме поле відведене для задання символу, який розділяє фонеми. Нарешті, передбачено секцію, в якій описуються можливі інтонації та відповідні їм символи.

Таку початкову інформацію задаємо як для окремої мови, так і для групи мов одночасно у випадку, якщо в поставленій задачі є елементи багатомовності.

Навчальною вибіркою називаємо сукупність наговорених диктором звукових файлів разом з відповідними орфографічними або фонетичними *текстами, які є окремими словами або фразами.* Звукові файли з навчальної вибірки ще називатимемо реалізаціями слова або фрази.

Накопичення навчальної вибірки за заданим текстом виконується за допомогою вікна накопичення навчальних вибірок. На початку задається ім'я (ідентифікатор) диктора та вводиться текст, орфографічний або фонетичний, за яким проводиться запис навчальної вибірки. Передбачено можливість запису реалізацій слів або фраз в окремі файли для зручності при подальшому обробленні даних.

Запис навчальної вибірки відбувається за такою схемою:

1. Мишкою активізується текст реалізації, з якої розпочинаємо запис.
2. Диктор натискає кнопку запису *Record* і, утримуючи її, промовляє відповідний текст.
3. Після промовляння тексту диктор відпускає кнопку запису.
4. Автоматично активізується наступна реалізація, якщо список не вичерпано. Вручну можна вибрати будь-який інший елемент.
5. Для запису наступної реалізації переходимо до кроку 2. Якщо записуваний текст вичерпано, завершуємо роботу.

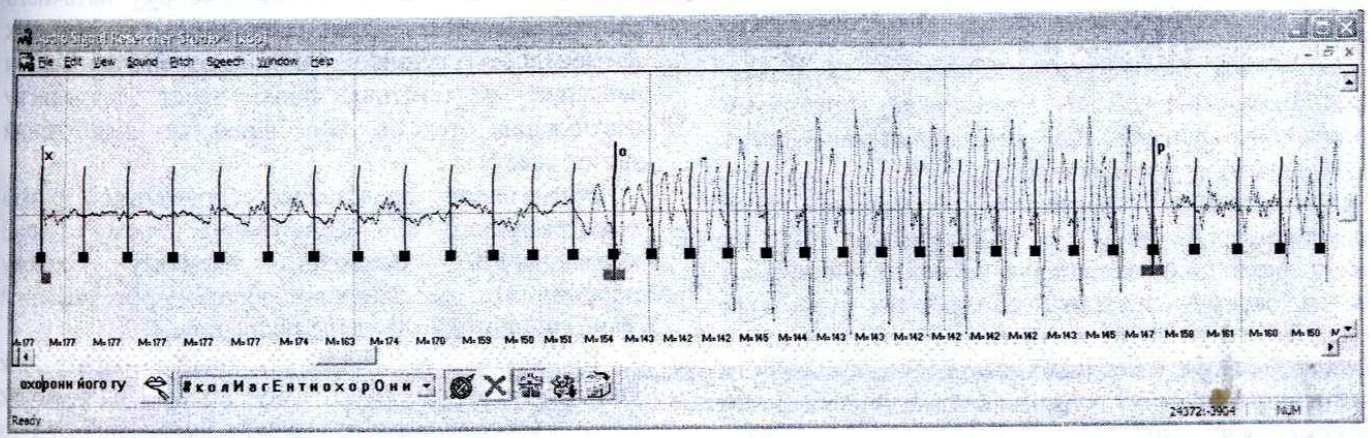

Рис. 1. Приклад зображення ділянки відсегментованої реалізації навчальної вибірки з чоловічого голосу. Сигнал розглядається крізь одноквазіперіодичну сітку, "натягнуту" як на періодичні сегмети, що відповідають фонемам-сонорантам, так і на неперіодичні сегменти, що відповідають глухим та паузі. Цифри знизу сигналу позначають тривалість квазіперіода в дискретах.

Записаний усномовний сигнал прослуховує експерт з метою виявлення бракованих або невиразних реалізацій, які позначаються як "погані" (*bad*). Остаточно, "погані" реалізації перезаписуються за вищенаведеною схемою з тією лише різницею, що автоматично активізуватиметься наступна відбракована реалізація.

В результаті процедури запису навчальної вибірки сформуються файли, що відповідають реалізаціям навчальної вибірки, яким співставлено ім'я диктора та текст, що було вимовлено. Далі проводимо сегментацію реалізацій навчальної вибірки на фонеми-трифони, а вибірки окремих фонем-трифонів – на одноквазіперіодичні мікросегменти та квазіперіодичні й неперіодичні сегменти. Причому, порядок сегментування може бути і зворотній.

На Рис. 1 показано приклад результату сегментування реалізації з навчальної вибірки. Ділянці усномовного сигналу співставлено відповідні фонеми-трифони, та відображено границі одноквазіперіодичних мікросегментів.

Сегментація навчальної вибірки на фонеми-трифони проводиться за участю експерта. Програмне забезпечення автоматично розставляє границі фонем лише дуже приблизно. Далі експерт, керуючись звуковою та візуальною підказкою уточнює межі фонем.

Пересування границь фонем здійснюється у двох режимах. Перший режим передбачає незалежне пересування границь сусідніх фонем, тобто можливі ділянки сигналу, на яких фонеми "накладаються" одна на одну, або ділянки, які не належать жодній фонемі. Коли активований другий режим, шляхом переведення кнопки *Linked Marks* у натиснутий стан, границі сусідніх фонем пересуваються узгоджено: початок поточної фонеми збігається із закінченням попередньої. В обох випадках за наявності одноквазіперіодичної розмітки границі фонем автоматично синхронізуються з границями квазіперіодів, тобто відбувається автоматичне прив'язування границь фонем-трифонів до одноквазіперіодичної сітки.

При уточнюванні границь фонем експерт використовує режим візуалізації сигналу, в якому зображується авторегресійний спектр, та режим прослуховування (*Sound->Sound Phone*), при якому озвучується фонема, на яку користувач-експерт вказує лівою кнопкою мишки.

Сегментація навчальної вибірки на одноквазіперіодичні мікросегменти та квазіперіодичні й неперіодичні сегменти виконується як на окремих звукових файлах, так і на всій навчальній вибірці в автоматичному режимі. Експерту лише необхідно вказати найбільшу та найменшу допустимі довжини квазіперіодів, а також обмеження на приріст тривалості одноквазіперіодів.

Границі квазіперіодів у вікні дослідження звукового сигналу зображаються у вигляді вертикальних рисок з кольоровими квадратиками-ручками в нижній частині риски. Ручки призначені для пересування границь квазіперіодів (мікрофонем).

Як тільки проведено сегментацію навчальної вибірки, інформація про фонеми-трифони та їх структуру вноситься до реєстру фонем-трифонів, формуючи таким чином усномовний файл диктора. Для цього слід натиснути кнопку *Update Collection*, попередньо виділивши сигнал тих фонем-трифонів, які заносяться до реєстру. Якщо не виділено жоної фонеми-трифона, до бази даних заносяться всі фонеми-трифони з реалізації.

Знання про індивідуальні інтонаційні контури отримуються шляхом оброблення відсегментованих реалізацій фраз, в яких представлені всі типи інтонації.

Сформований усномовний файл диктора зберігається у вигляді файлу, який містить не сам сигнал, а лише посилання на фрагменти звукового файлу, що відповідають фонемам-трифонам. Таким чином, зберігається зв'язок з навчальною вибіркою, і зміни сегментації, проведені на навчальній вибірці, автоматично відображаються в базу даних шляхом виклику команди *Update Collection*.

4 Дослідження синтезованого сигналу

В рамках Студії дослідника усномовного сигналу розроблено стенди дослідження синтезу та розпізнавання усного мовлення [5]. З використанням першого стенду виконуються експериментальні дослідження розбірливості та якості звучання синтезованого усномовного сигналу. На другому стенді проводяться дослідження пофонемного розпізнавання.

Стенд дослідження усномовного синтезу являє собою вікно, представлене діалоговою панеллю з багатьма елементами регулювання та командними елементами, які логічно розбиті на частини, як це показано на Рис. 2.

Поле задання орфографічного або фонетичного тексту, що буде подано на озвучення, дозволяє маніпулювати зі зразками тексту або транскрипції, озвучений сигнал яких планується дослідити. Це поле складається з вікна вибору поточного озвучуваного тексту, текстового вікна та командних елементів – гудзиків, що спонукають додати введений у текстове вікно текст до списку озвучуваних текстів або видалити непотрібний зразок тексту.

Вікно вибору представляє собою список зразків орфографічного або фонетичного тексту. Перші кільканадцять символів елементу списку допомагають дослідникові обрати орфографічний /фонетичний текст із уведених раніше.

Вибраний зі списку текст одразу ж відображається у текстовому вікні. Це дозволяє бачити текст цілком, редагувати його. Щоб додати новий зразок тексту, орфографічного або фонетичного, слід увести зразок тексту до текстового вікна або відредагувати текст попередньо доданого зразка і виконати команду додання нового зразка

тексту шляхом активації відповідного командного гудзика. Також передбачена можливість видалити зразок тексту.

Щойно додано новий зразок тексту або вибрано з уведених раніше зразків, поле виділення фонем-трифонів відображає фонемно-трифонні транскрипції, що відповідають поточному зразкові тексту. У поточній фонемно-трифонній транскрипції передбачено можливість виділяти окремі фонеми-трифони, які з метою досліджень можна замінювати на інші та/або задавати їм різні просодичні характеристики за допомогою описаних далі полів.

Поле вибору фонем-трифонів дозволяє "підсаджувати" різні фонеми-трифони з індивідуального усномовного файлу диктора у фонемно-трифонну послідовність озвучуваного тексту.

Основний елемент поля вибору фонем-трифонів є таблиця фонем-трифонів, доступних у індивідуальному усномовному файлі диктора, що містить база даних і знань озвучення текстів. Таблиця показує лише фонеми-трифони-претенденти, тобто ті фонеми-трифони, чия термінальна фонема збігається з термінальною фонемою виділеної фонеми-трифона у полі фонемно-трифонної транскрипції.

У таблиці коротко характеризуються фонеми-трифони-претенденти, показано їх оригінальне фонемне оточення. Щоб замінити виділену фонему-трифон у поточній фонемно-трифонній послідовності озвучуваного тексту, необхідно вибрати фонему-трифон з таблиці та викликати команду *Assign*.

Поле регулювання просодичних характеристик сигналу дозволяє змінювати інтонацію, темп і гучність синтезованого сигналу. Це поле містить список всіх одноквазіперіодичних сегментів (мікросегментів) виділеної фонеми-трифона поточної фонемно-трифонної транскрипції. У сусідніх текстових віконечках відображаються значення довжини вибраного одноквазіперіоду та його гучності. Передбачена можливість розмножувати окремі квазіперіоди.

Шляхом зміни довжини квазіперіодів фонеми-трифона в допустимих межах відбувається зміна інтонаційного контуру фонеми на низькому рівні. Викидання окремих мікросегментів або їх множення призводить відповідно до скорочення або подовження фонеми-тифона вцілому, а отже до зміни його темпоральних характеристик.

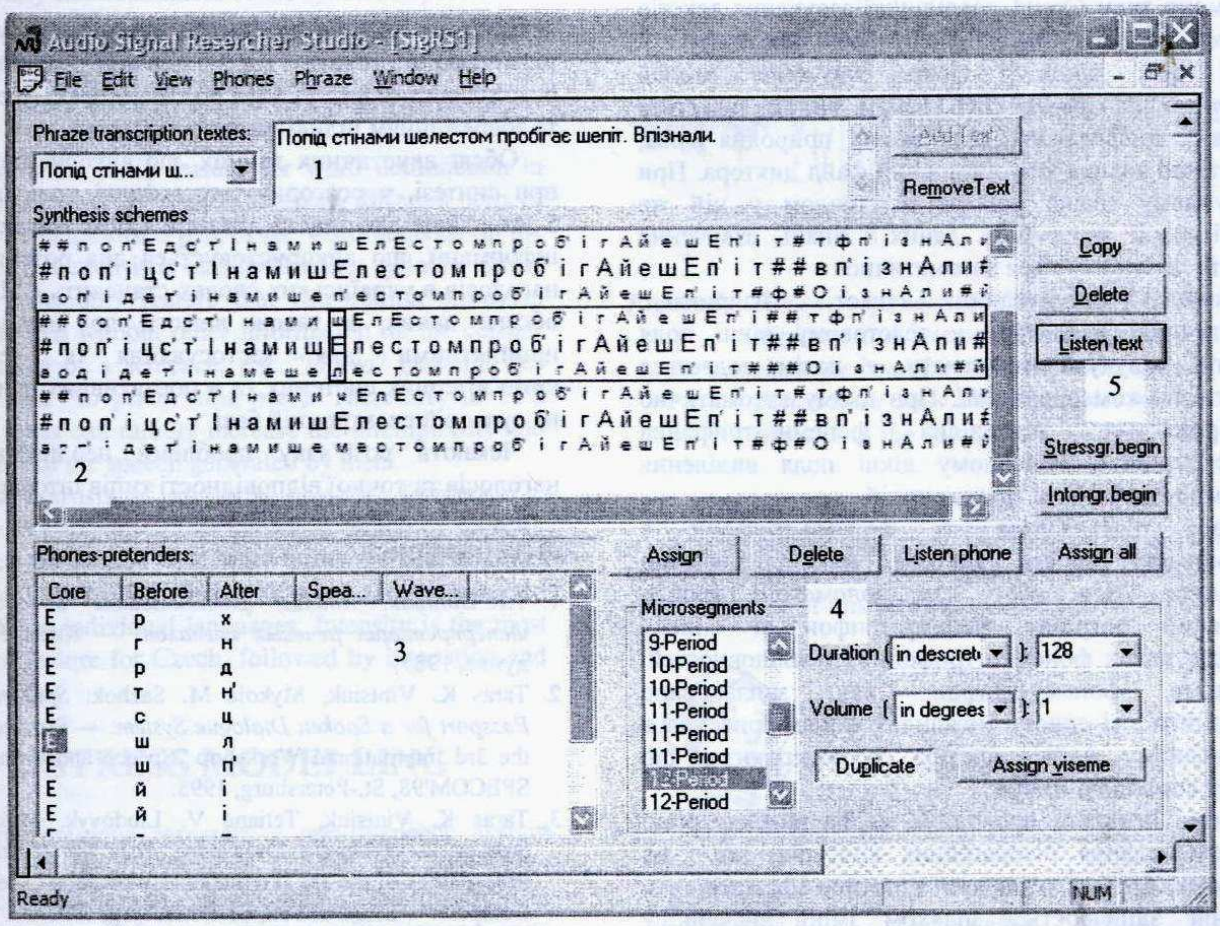

Рис. 2. Вікно дослідження усномовного синтезу представляє собою діалогову панель з елементами регулювання та командними елементами, які підрозділяються на такі розділи:

1) секція задання озвучуваного тексту, орфографічного або фонетичного;

2) інтерактивний список фонемно-трифонних транскрипцій заданого орфографічного (фонетичного) тексту;

3) секція вибору фонем-трифонів з бази даних і знань;

4) секція регулювання просодичних характеристик сигналу;

5) загальні командні елементи.

Передбачена також можливість зміни інтонаційних харктеристик на вищому рівні шляхом задання інтонаційних контурів синтезованого тексту. Відповідна команда у меню (*Phraze->Apply intonation*) викликає діалогове вікно, в якому за спеціальною схемою описуються інтонаційні контури ритмогрупи або синтагматичні. Після введення опису інтонаційних контурів і підтвердження операції, просодичні параметри керування синтезатором можуть бути обчислені за одним з уведених контурів.

Загальні командні елементи містяться як на стенді, окремо від описаних вище секцій, так і в меню.

Після заповнення належним чином полів стенду досліджень усномовного сигналу подається загальна команда озвучення тексту. На самому початку роботи зі стендом необхідно завантажити базу даних і знань для озвучення українських текстів загальною командою з меню. Одночасно можливо працювати як з декількома базами даних, так і з різними стендами дослідження синтезу усної мови.

Загальний порядок роботи зі стендом дослідника усномовного синтезу виглядає наступним чином. Щоб започаткувати нові дослідження, у Студії дослідника усномовного сигналу створюємо новий документ типу Стенд дослідника озвучення текстів (*Speech Synthesis Master*). Далі необхідно викликати файл бази даних і знань озвучення текстів (*Phones->Add speech DB*). Таким чином задається основна конфігурація досліджень: природня мова, фонетичні знання та усномовний файл диктора. При наступному сеансі роботи зі стендом у цій же конфігурації файл бази даних і знань озвучення текстів завантажується автоматично.

Новий озвучуваний текст, наприклад, фонетичний, вводиться у текстовому вікні поля задання озвучуваного тексту, а потім додається супутньою командою *Add*. При цьому автоматично відображається відповідна фонемно-трифонна транскрипція у головному вікні поля виділення фонемно-трифонних транскрипцій.

Далі підсаджуємо ті фонеми-трифони з усномовного файлу диктора, які заплановано дослідити. Для цього за допомогою мишки виділяємо потрібну фонему-трифон зі списку досліджуваних фонемно-трифонних послідовностей, обираємо фонему-трифон, яку заплановано "підсадити", зі списку можливих фонем-трифонів в усномовному файлі диктора, і закріплюємо цю заміну командою *Assign*.

Зміни інтонації проводимо як на вищому рівні шляхом задання інтонаційних контурів, так і на нижчому, вручну змінюючи довжини квазіперіодів. Останнє здійснюється шляхом зміни значень у віконечку довжини виділеного квазіперіоду. Щойно точка введення перейде до іншого елемента стенду (мишкою виділяємо наступний квазіперіод), змінені значення довжини квазіперіоду запам'ятаються і будуть використані при наступній процедурі синтезу сигналу.

Темпоральні зміни прототипу на низькому рівні здійснюються за допомогою маніпуляцій з мікросегментами фонем-трифонів. Множення виділених мікросегментів за допомогою команди *Duplicate* призводить до подовження окремої фонеми-трифона, а отже і до загального сповільнення темпу. Викидання мікросегментів, тобто задання їм нульової довжини, призводить до скорочення довжини фонеми-трифона, а тому – до прискорення темпу.

Нарешті, командою *Listen text* запускаємо процедуру синтезу, яка моделює оригінальний алгоритм синтезу усномовного сигналу в часово-амплітудній області. При цьому відкривається новий документ типу *Усномовний сигнал*, куди автоматично вставляється синтезований сигнал, який тепер є доступним для прослуховування, аналізу та подальшого дослідження (Рис. 1).

При прослуховуванні експертами синтезованих як окремих слів, так і злитого мовлення, було з'ясовано словесну розбірливість синтезованого сигналу, що виявилася на рівні не менше 90% на перших сенсах, зростаючи за мірою "звикання" до синтезованого голосу.

5 Висновки

Запропонований метод синтезу дозволяє озвучувати українські тексти з доволі прийнятною розбірливістю, натуральністю синтезованого сигналу та збереженням індивідуальності мовця.

Обсяг акустичних данних, що використовуються при синтезі, у розгорнутому вигляді становить від 5 МБ і вище для одного диктора. Обсяг лінгвістичної інформації, що використовується для розставлення наголосів в українських словах становить 4 МБ. Такі обсяги даних, а також вимоги до швидкодії є прийнятними для застосування в реальних комп'ютерних системах та в портативних пристроях на сучасній електронній базі.

Чекають розв'язку проблеми неоднозначності наголосів та точної відповідності типів інтонації.

Література

1. Т.К. Винцюк. *Анализ, распознавание и смысловая интерпретация речевых сигналов.* — Киев: Наукова думка, 1987.
2. Taras K. Vintsiuk, Mykola M. Sazhok: *Speaker Voice Passport for a Spoken Dialogue System.* — Proceedings of the 3rd International Workshop "Speech and Computer" - SPECOM'98, St.-Petersburg, 1998.
3. Taras K. Vintsiuk, Tetiana V. Liudovyk, Mykola M. Sazhok. — *Phonetic Knowledge Base for Ukrainian.*— Proceedings of the 3rd International Workshop "Speech and Computer" - SPECOM'98, St.-Petersburg, 1998.
4. T. Dutoit, H. Leich. — A Comparison of Four Candidate Algorithms in the Context of High Quality Text to Speech Synthesis. — ICASSP'94.
5. Микола Сажок. *Комп'ютерні засоби експериментальних досліджень усномовного сигналу.* — Праці 4-ї Всеукраїнської міжнародної конференції "УкрОбраз-98", Київ, 1998.

Modelling Word Stress for Use in Speech Synthesis

Daniel Tihelka, Jindřich Matoušek, Martin Vlach

University of West Bohemia in Pilsen, Department of Cybernetics,
Univerzitní 22, 306 14 Pilsen, Czech Republic
dtihelka@kky.zcu.cz, jmatouse@kky.zcu.cz

ABSTRACT

This paper deals with word stress modelling for use in text-to-speech systems. It describes one of the ways of stress modelling. Several experiments were carried out for different phonemes containing stress and other prosodic features. Listening tests with specially designed words were set up for the comparison of speech generated from each experiment. The results show that the examined way of stress modelling can increase intelligibility and naturalness of the synthetic speech.

1. INTRODUCTION

Stress is the basic feature for word delimitation in continuous spoken speech, emphasising different parts of words. The emphasised part of the word depends on the language; in Czech it is usually the first syllable. Stress is mostly the only feature that can delimit words in spoken speech without other information on the context and meaning of a sentence, e.g. "*topivo*" (material for burning) and "*to pivo*" (this beer). Therefore, in TTS systems stress can rapidly increase the intelligibility and naturalness of the speech generated by them.

Stress consists of all three prosodic features, i.e. *duration* of phonemes (also known as tempo of speech), *intonation* (frequency F_0 of voice base tone) and *intensity* (loudness). Each of these features is emphasised differently in individual languages. Intensity is the most important feature for Czech, followed by intonation and duration [3].

2. STRESS MODELLING

In our previous research, a text-to-speech system ARTIC (ARtificial Talker in Czech) was built. This system is based on unit concatenation in the time domain [2] and was used for experiments with stressed units presented in this paper.

A new speech corpus recorded by a female speaker, comprising several hours of speech, was used for building the speech segment database (SSD) for this system. A fully automatic process based on hidden Markov models (HMM) was used in a segmentation process, and the speech corpus was segmented into *fenemes* (each feneme corresponds to one state of HMM; each HMM is used for the modelling of a triphone, see Figure 1) [2].

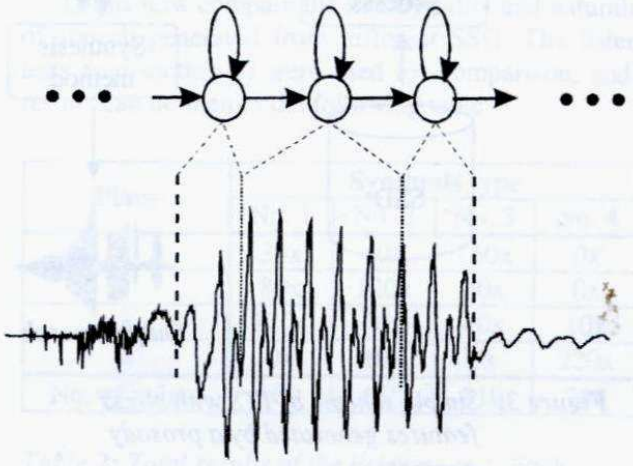

Figure 1: *The correspondence of three-state HMM, a triphone and fenemes. The dashed lines show the boundary of a triphone, the dotted lines show the boundaries of the fenemes of the vowel "a".*

Representative segments were stored in the speech segment database after the segmentation process. A very simple scheme of this process can bee seen in Figure 2.

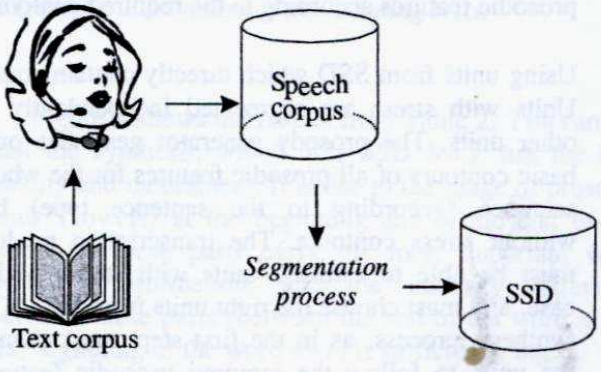

Figure 2: *Simply scheme of segmentation process*

ARTIC can work with prosodic features. All these are represented as a contour of changes from the base value, and the concatenation method can modify concatenated units according to contour requirements.

There are two basic ways of modelling the stress features in TTS systems:

- Using the contours from prosodic features for stress modelling, especially the contour of intensity and intonation for Czech speech. For a simple scheme of this system see Figure 3. In this case, the prosody

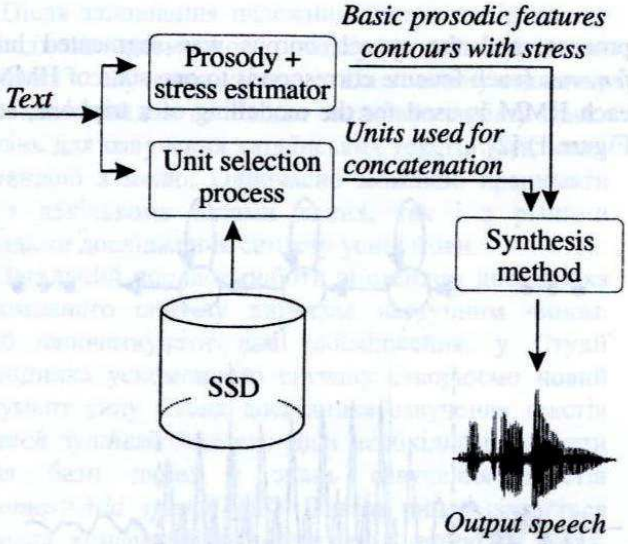

Figure 3: *Simple scheme of TTS with stress features generated by a prosody generator.*

generator estimates the basic contours of all prosodic features for the whole sentence (these features depend on the sentence type). The stress features of all stressed phonemes must be added to the basic contours. This means that the generator must be able to estimate phonemes with stress. The transcription module provides transcription to the concatenated units (fenemes in our case) stored in SSD, which are independent of stress. The synthesis module joins units together and changes their prosodic features according to the required contours.

- Using units from SSD which directly contain stress. Units with stress are segmented independently of other units. The prosody generator generates only basic contours of all prosodic features for the whole sentence (according to the sentence type) but without stress contours. The transcription module must be able to estimate units with stress in this case, and must choose the right units from SSD. The synthesis process, as in the first step, must change the units to follow the required prosodic features contours. A simple system scheme can be seen in Figure 4. In our research we focused on the second approach, as described below.

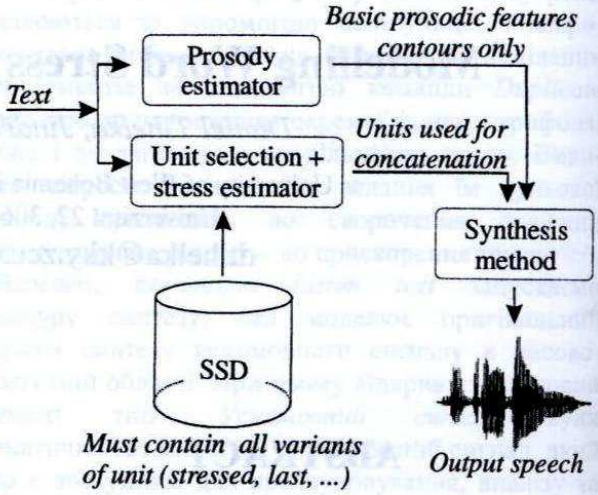

Figure 4: *Simple scheme of TTS using units with stress features*

3. PHONEMES USED AS PROSODIC UNITS

The statistical approach based on hidden Markov models was used, as mentioned above. The following text describes the positions and types of phonemes chosen for prosody modelling:

1. At first, the vowel in the first syllable of each word with one or more syllables was modelled as *stressed,* which means that this unit contains stress. All other vowels and consonants were modelled as unstressed, because, as we mentioned, stress lies on the first syllable in Czech, and intensity change is usually bigger in a vowel than in a consonant. It can be expected that this simple approach will increase the intelligibility of speech without a large increase in the size of SSD.

2. The vowel in the first syllable (as in the previous case) was modelled as stressed, and the vowel in the last syllable in the words with two and more syllables was modelled as *final.* This means that it lies at the end of the word, and it is possible that the next syllable will be stressed. All the other units were modelled as unstressed. A bigger intensity and intonation contrast between the stressed syllable and the final syllable can be expected, increasing the delimiting feature of stress in Czech. Generated speech could be more intelligible and the size of SSD could still be reasonable.

3. All vowels and consonants in the first syllable were modelled as stressed and all vowels and consonants in the last syllable were modelled as final. All other triphones (i.e. triphones in the middle syllables) were modelled as unstressed. Syllables are modelled

with a greater precision in this approach, as every triphone of the word contains information about its position in the word. We expected that this approach would bring the highest intelligibility and naturalness. Using this approach, the size of SSD will be the largest.

Three-state left-to-right models were used for all described units. These models had the same initialisation of the state mean vectors, covariance matrices and transition matrices before the training process.

3. THE METHOD FOR RESULTS COMPARISON

We set up small listening tests for the comparison of results obtained with different prosodic units. These tests were designed especially for Czech phenomena on words boundaries.

Here are some examples of word pairs which are differentiated from each other by stress positions. The individual words were used in the sentences for the listening tests:

topivo	(fuel)	*to pivo*	(that beer)
tabulka	(a chart)	*ta bulka*	(that roll)
prsten	(a ring)	*prs ten*	(that breast)
jak oběžné	(as circular)	*jako běžné*	(as usual)

Etc.

There were 15 sentences containing one of these words. Each of these 15 sentences was synthesised from:

- SSD without stressed units at first (synthesis output No.1 in the following tables).
- SSD containing units for vowels in the first syllable of the word (synthesis output No.2).
- SSD with units for vowels in the first and in the last syllables of the word (synthesis output No.3).
- SSD with units for all phonemes in the first and in the last syllables of the word (synthesis output No.4).

Naturally, for each synthesis output mentioned, ARTIC had to be able to select the right units for each used SSD.

23 people took part in the listening tests (14 female and 9 male listeners). Every listener heard 10 sentences in all their versions (4 versions per sentence) and had to select the sequence of types of each sentence from the best to the worst. The best type was evaluated by 4 points, the worst by 1 point.

4. RESULTS

Let us compare the sizes of the Speech Segment Database first.

As can bee seen in Table 1, the size of SSD was very similar for syntheses No. 1 and 2, as synthesis No.2 has only 1011 stressed units (triphones, corresponding to vowels at the beginning of the word). The size of SSD for synthesis No.3 is still not very big, as 527 units were added for modelling vowels at the end of the word. On the other hand, the size of SSD used for synthesis No.4 is a little bigger.

SSD used	Number of units in SSD	Size of SSD [MB]
No. 1	9097	17.501
No. 2	10108	20.808
No. 3	10635	31.165
No. 4	11877	47.566

***Table.1** SSD sizes for different prosodic units and number of units in the SSD.*

Let us now compare the intelligibility and naturalness of speech generated from different SSD. The listening tests (see section 3) were used for comparison, and the results can be seen in the following table.

Place	Synthesis type			
	No. 1	No. 2	No. 3	No. 4
1	30x	40x	160x	0x
2	80x	120x	30x	0x
3	120x	60x	40x	10x
4	0x	10x	0x	220x
No. of points	600	650	810	240

***Table 2:** Total results of the listening test. Each number represents the number of occurrences at a given place.*

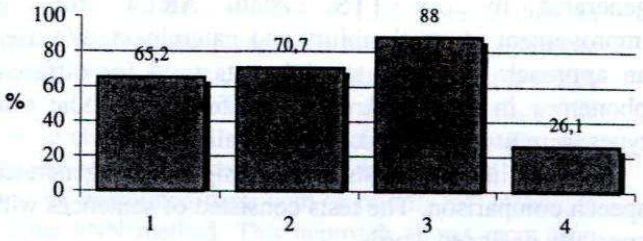

***Figure 5:** Evaluation of the listening tests.*

Figure 5 shows the results from Table 2. You can see that the synthesis, which uses SSD No.3, has the best quality and naturalness. It is due to the usage of prosodic units (vowels) at the beginning and at the end of the word, as these parts carry the most important word prosodic information, and the intensity difference between these parts (between the end of the word *n* and the beginning of the word *n+1*) is sufficiently big.

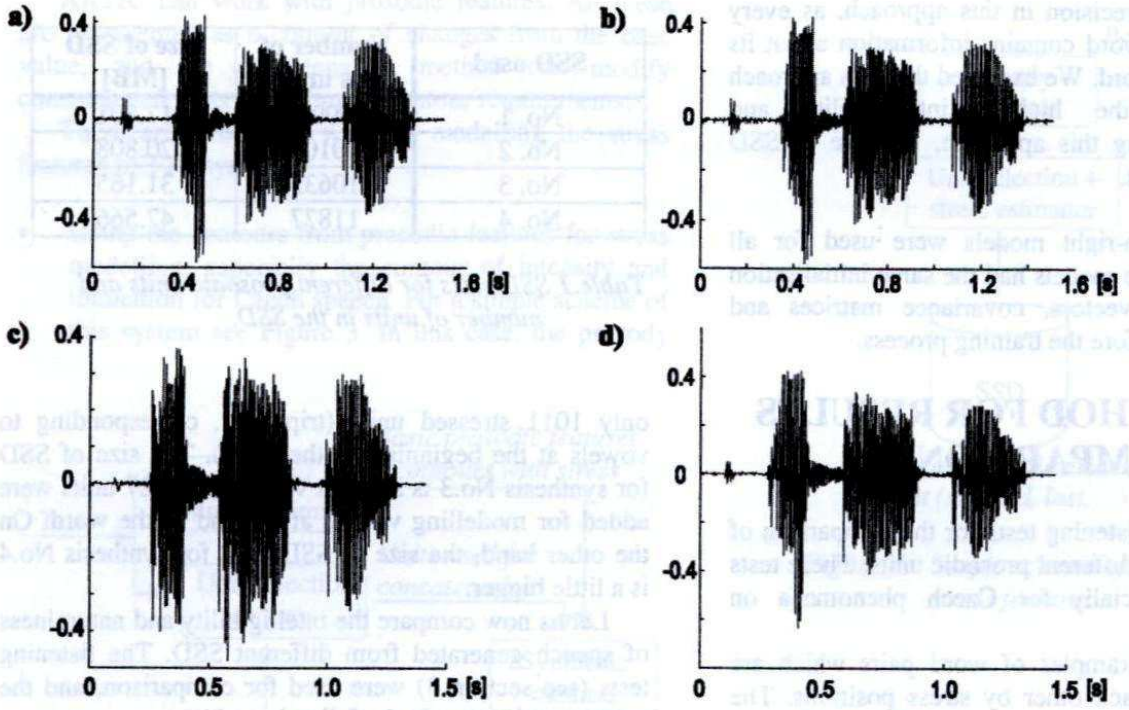

Figure 6: Speech signals for the words "tabulka" (a, c) and "ta bulka" (b, d) from SSD without prosodic units (a, b) and with stressed and final vowels as prosodic units (c, d).

Using SSD No.4 for speech generation, the best results were expected. However, this speech turned out to be the worst. It is due to the small number of units used for training the HMMs, as every unit from SSD No.1 had 3 variants in the SSD No.4.

7. CONCLUSION

We tried to add the stress of the words to the speech generated by our TTS system ARTIC for the improvement of intelligibility and naturalness. We used an approach based on special units used for different phonemes in the word. Three different prosodic unit types were used: stressed, final and all others.

Special listening tests were designed for generated speech comparison. The tests consisted of sentences with specially designed words.

The listening tests showed that this approach brought higher intelligibility and naturalness, especially when vowels in stressed and final syllables of the word were modelled independently of other phonemes.

Stress modelling directly by a prosody generator is planned as our future work.

AKNOWLEDGEMENT

This research has been supported by the Grant Agency of the Czech Republic No. 102/02/P134 and by the Ministry of Education of the Czech Republic No. MSM 235200004.

REFERENCES

1. Matoušek J., Psutka J. and Krůta J.: "*Design of Speech Corpus for Text-to-Speech Synthesis.*" -In: Proceedings of the EUROSPEECH 2001, vol. 3. Aalborg, Denmark, 2001, pp. 2047-2050.

2. Matoušek J. and Psutka J.: "*ARTIC: A New Czech Text-to-Speech System Using Statistical Approach to Speech Segment Database Construction.*" -In: The Proceedings of ICSLP2000, vol. IV. Beijing, China, 2000, pp. 612-615.

3. Palková Z.: "*Fonetika a fonologie češtiny (Phonetics and phonology of Czech).*", Karolinum, Prague 1994.

4. Donovan R.E., Woodland P.C.: "*A Hidden Markov-Model-Based Trainable Speech Synthesiser.*", Computer Speech and Language, 1999.

Automatic phonetic baseforms for person names in the Czech dialogue system

Pavel Brodský, Luděk Müller
University of West Bohemia in Pilsen, Department of Cybernetics,
Univerzitní 22, 306 14 Pilsen, Czech Republic
pbrodsky@kky.zcu.cz, muller@kky.zcu.cz

ABSTRACT

Phonetic baseforms are the basic recognition units in most speech recognition systems. These baseforms are usually determined by linguists once a vocabulary is chosen and not modified thereafter. However, several applications, such as name dialling require the user to be able to add new words to the vocabulary. These new words are often names or task-specific jargons that have user dependent pronunciation.

In this paper, we describe a method how to generate a phonetic baseform from acoustic pronunciation of a name without a prior knowledge of the name spelling. We used a language model based on bigram statistics.

1. INTRODUCTION

There has been considerable interest in *telecommunications and embedded speech recognition* application that provide personalized vocabularies. Name dialling is one such example of a telephonic application where it is necessary to have ability to provide speaker dependent vocabularies for repertory dialling. This feature enables the user to add even such words to the personalized vocabulary for which a spelling or acoustic representation does not exist in the speech recognition lexicon, and associate these words to a phone number to be dialled. We will show how speaker dependent baseforms could be derived from one or two speech utterances by using speaker independent acoustic model and a language model. We use the bigram probabilities to constrain the transition between phonemes.

The structure of this paper is as follows. In Section 2 we present our recognition system and its components: Speech recognition engine, acoustic modelling, front-end, labeller and decoder. In Section 3 we describe baseform generating algorithm. The mumble model is described in Section 4. In Section 5 language model for names and surnames of Czech Republic inhabitants is described. Experimental results are contained in Section 6 and Section 7 is conclusion.

2. SYSTEM OVERVIEW

The speech recognition engine is based on a statistical approach. It comprises a front-end, an acoustic model, a language model and a decoding block [1].

Acoustic Modelling: As a basic speech unit of the recognition system a triphone is used. Each triphone is represented by a 3-state left-to-right HMM with a continuous output probability density function assigned to each state. Each density is expressed as a mixture of multivariate Gaussians with a diagonal covariance matrix. The Czech phonetic decision trees were used to tie states of Czech triphones.

Front-end: The speech signal is digitized through a telephone board at 8 kHz sample rate and converted to the mu-law 8-bit resolution format. The parametrization process used in our system is as follows: Firstly the pre-emphasized acoustic waveform is segmented into 25 millisecond frames every 10ms. A Hamming window is applied to each frame and 13 MFCCs (including the energy coefficient c_0) are computed. The first-order and second-order derivates of MFCCs are computed and appended to the static MFCCs each speech frame.

Labeller: The recognition algorithm uses 2510 different tie states, each of which represented by a mixture of 8 Gaussian distributions in the 39-dimensional space. Thus during a decoding it is necessary to compute a large number of log-likelihood scores (LLSs) every 10ms. In order to perform the recognition in real time the number of calculations is reduced by applying a technique which seeks to find and precisely determinate only first 150 most probable LLSs. This technique efficiently uses relevant statistical properties of the Gaussian mixture densities combining them with "a priority hit" technique and the kNN method. This approach allows more than 90% reduction of a computation cost without substantial decrease recognition accuracy.

Decoder: The decoder uses a crossword context dependent HMM state network generated by a Net generator. The input of the Net generator is a text grammar format represented by an extended BNF with respect of JSGF. The whole net consists of one or more in run-time connected regular grammars. A considerable part of the net is usually generated before the decoder starts but every part of the net can be generated on

demand in run-time. The decoder utilizes a Viterbi search with a beam pruning.

3. PHONETIC BASEFORMS

Phonetic baseforms are the basic recognition units in most speech recognition systems. These baseforms are usually determined by linguists once a vocabulary is chosen and not modified thereafter. However, several applications, such as name dialling, require the user be able to add new words to the vocabulary. These new words are often names or task-specific jargons that have user dependent pronunciations.

The phonetic baseform is sequence of phonemes which represents a given utterance (name). We can create it manually (by phonetic transcription rules) or automatically.

Example of baseform (for the Czech phonetic alphabet):

Name:	Luděk Müller
Manually:	sil l u dj e k sil m i l e r sil
Automatic:	sil r i d e p t sil m e l aa r sil

Some utterances can be pronounced by several ways depending on speaker, speaking style, and other conditions. Therefore, generally more than one baseform can be constructed and stored for an utterance (e.g. a person name). In this work we used only one phonetic baseform for each person name.

Each baseform can be easily added manually. In the case of automatic baseform generation the user for example can be required to say the given person names twice and for each utterance a baseform should be automatically generated.

The baseform generation algorithm is based on Viterbi algorithm [3] that searches the best phoneme path through the HMM net consist of all Czech phonemes and a set of phoneme transitions. The net structure is dependent on the language (phoneme) model and can be interpreted also as so-called mumble model described in more detail in the Section 4.

Also N-best hypotheses instead the first best hypothesis can be considered and in this case the decoder should produce a list of the most probable phoneme sequences. The recognition results can be stored as well in a phoneme graph which is an analogy to the word graph in the case of N-best word sequences decoding problem.

4. MUMBLE MODEL

The mumble model is constructed as a set of HMM [4] models connected in a parallel fashion. Each HMM model is 3-state left-to-right and represents one context-independent phone. The structure of the mumble model is depicted in Figure 1. Actually the probabilities of emission of an observation vector in a given state are evaluated as the maximal emission probability of all corresponding states of context-dependent triphones. Thus neither additional HMM models nor additional training is required. The value of the backward loop probability *BPr* causes a various length of the phone sequence recognized by the network in Figure 1. While the higher value produces more insertions, the small value induces more deletions in the resulting phone sequence.

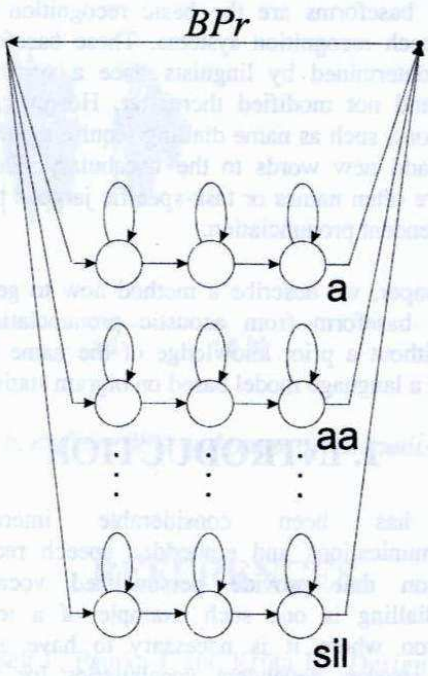

Figure 1.

In Figure 2. is mumble model with language model basis on bigrams with full matrix of transition probabilities. Language model is described in Section 5.

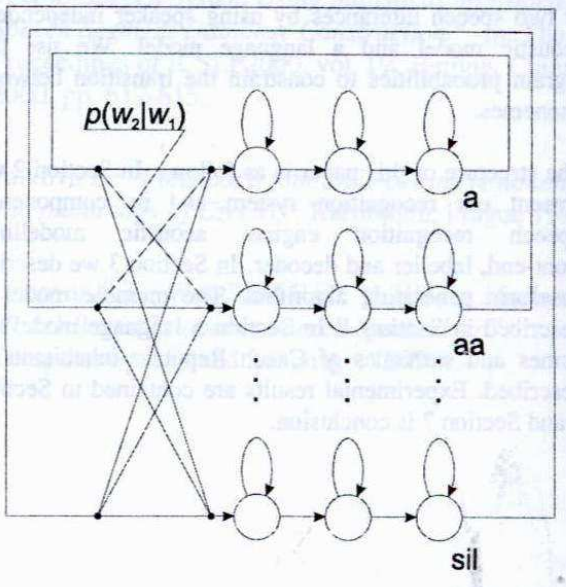

Figure 2.

5. LANGUAGE MODEL

The language model generally restricts variety of words sequences W and consequently also phone sequences. This restriction may be either deterministic (i. e. some words or phones sequence are not allowable) or "softer" stochastic (some word sequence are less probable). In our case we chose a probabilistic approach and an absolute discounting language model with backing-off for conditional probabilities method.

The basic idea of the absolute discounting language model with backing-off for conditional probabilities is to keep a high number of joined events (h, w), a word history h and a word w, almost unmodified. We suppose that the number of occurrences of joined events in training text will not change probably too much, if we select another training text of the similar size (from the same problem area). To consider possible variability of the number $N(h, w)$ of occurrences (h, w) in text we introduce parameter of permanent deviation b_h, so-called absolute discounting parameter that decreases the number of seen events $N(h, w)$. Furthermore we suppose that b_h is not dependent directly on the value $N(h, w)$, nevertheless it is dependent on the history h. This deviation must remain negative because unseen events in the text requires nonzero (thus positive) probabilities. By means of absolute discounting parameter b_h the part of probability mass is redistributed from the seen events to unseen events. The resulted formulae for the absolute discounting language model is:

$$\bar{p}(w|h) = \begin{cases} \dfrac{N(h,w) - b_h}{N(h,.)}, & \text{for } N(h,w) > 0 \\ b_h \dfrac{V - n_0(h,.)}{N(h,.)} \cdot \dfrac{\beta(w|\bar{h})}{\sum_{w':N(h,w')=0} \beta(w'|\bar{h})}, & \text{for } N(h,w) = 0 \end{cases} \quad (1)$$

where $\beta(w|\bar{h})$ is a conditional probability of observing the word w given the generalized history $\bar{h}$. The generalized (also reduced) history $\bar{h}$ is defined as:

If (h,w) n-gram $(w_1, w_2, ..., w_n)$, then $(\bar{h}, w)$ is $(n\text{-}1)$-gram $(w_2, ..., w_n)$.

Our language model was created for names and surnames occurring in Czech Republic. We had at disposal 4 137 different names and 236 769 different surnames. In total we had 10 282 470 names and 10 296 459 surnames. As a basic unit of the language model we chose a phoneme which means that before computing individual probabilities we had to perform a phonetic transcription. Foreign names and surnames that are not subjected to Czech phonetic transcriptions rules were transcribed manually and saved into vocabulary of exceptions. After the phonetic transcription we computed probability of unigrams, bigrams and trigrams according to equation (1). Results are shown in Table 1. including the task perplexity and entropy.

Characteristics of training and test corpus			
		training	test
number of names		20 579 070	9 488
number of phonemes in vocabulary (V)		44	44
unigrams	unigrams (N)	174 692 140	77 100
	different unigrams (nr(.,.))	44	43
	unseen unigrams (n0(.,.))	0	1
	singletons (n1(.,.))	0	0
	doubletons (n2(.,.))	0	0
	perplexity	19.22	19.03
	entropy	4.26	4.25
bigrams	bigrams (N)	154 112 095	67 612
	different bigrams (nr(.,.))	1 392	852
	unseen bigrams (n0(.,.))	544	1 084
	singletons (n1(.,.))	35	101
	doubletons (n2(.,.))	13	60
	perplexity	9.95	10.11
	entropy	3.32	3.34
trigrams	trigrams (N)	133 532 050	58 124
	different trigrams (nr(.,.))	18 904	4 999
	unseen trigrams (n0(.,.))	66 280	80 185
	singletons (n1(.,.))	1 198	1 725
	doubletons (n2(.,.))	645	757
	perplexity	4.70	4.54
	entropy	2.23	2.18

Table 1.

6. EXPERIMENTAL RESULT

In the speech recognition system equipped by an acoustic and a language model it is practically advantageous to use different weights of the language and the acoustic model. These weights can be defined by two variables p a s. The word insertion penalty p is a fixed value added to each token when it transits from the end of one word to the start of the next. The grammar scale factor s is the amount by which the language model probability is scaled before being added to each hypothesis as it transits from the end of the word to the start of the next. These parameters can have a significant effect on recognition performance and hence, some tuning on development test data is well worthwhile.
Formulae for computing with parameters s and p:

$$\log(P(O, w_k|w_l)) = \log(P(O|w_k)) + s \cdot \log(P(w_k|w_l)) + p \quad (2)$$

where $\boldsymbol{O}$ is observation vector sequence generated by Hidden Markov Model; w_k and w_l are phonemes and P is a likelihood.

After the language model had been created we performed several tests. The first test was performed using 718 utterances. Each utterance consists of person name and surname. From 718 utterances two sets A, B were randomly chosen. Each of them contained 25 different utterances. Two training sets were constructed. The first one (T1) is equal to the set A and the second one (T2) is composes both sets A and B. The test set contained all 718 utterances. From sets A and B the baseforms were generated. The test set was recognized on basis these baseforms. Results are shown in Table 2. for the grammar scale factor $s = 1$ and the word insertion penalty $p = 30\,000$.

Number of baseforms	25 (T1)	2x25 (T2)
Number of utterances	718	718
Correctly recognized	554	601
Incorrectly recognized	164	117
Correctly recognized [%]	77.16	83.70
Incorrectly recognized [%]	22.84	16.30

Table 2.

Following series of tests were performed with the same set of 718 utterances. 25 utterances were always randomly chosen (one for every name) and all the 718 utterances were subsequently recognized with various values s and p. General results shown in Table 3. are arithmetic mean of all individual tests.

s/p	20000	25000	30000	35000	40000	50000
-1	75.81	75.07	72.28	68.71	64.02	59.52
0	79.53	79.20	75.91	74.74	70.98	63.37
1	79.02	78.46	77.21	76.42	74.47	68.48
2	76.04	78.55	79.06	77.11	76.93	74.65

Table 3.

In the final test we tried how the system works for one user. We recorded 200 utterances by one user (4 utterances for each full name from the test A. From 200 utterances we randomly chose 25 utterances of different names and for each utterance a baseform was generated. The rest (175) utterances were recognized with these baseforms. We executed two tests with different training utterances. Results shown in Table 4. are arithmetic mean of both the tests.

s/p	20000	25000	30000	35000	40000	50000
0	94	93	92	90,5	88	84
1	93.5	92.5	93	95.5	91.5	87.5
2	99.5	97.5	91.5	89	89	90.5

Table 4.

7. CONCLUSION

We have presented a system for recognition and automatic phonetic baseform generation. We used an acoustic model and a language model with bigram probabilities to constrain the transitions between phonemes.

In conclusion, we believe that we have a viable technique for automatic generation of phonetic baseforms that give a good decoding accuracy with our speech recognition system. This is particularly useful for our telephony dialogue system where personalized vocabularies are a must.

Experimental results show that the recognition based on acoustic baseforms has a good accuracy. The highest achieved accuracy ($s = 2$, $p = 20\,000$) for system working with one user is greater then 99 %.

In the future we want to provide further improvements in accuracy.

REFERENCES

[1] Müller L., Psutka J., and Šmídl L. "*Design of Speech Recognition Engine*", TSD 2000 – Third International Workshop on TEXT, SPEECH and DIALOGUE, Brno, Czech republic, September 13-16, 2000.

[2] Ramabhadran B., Bahl L. R., deSouza P. V., Padmanabhan M., "*Acoustic-Only Based Automatic Phonetic Baseform Generation*", ICASSP 1998

[3] Forney, G. D., Jr. „*The Viterbi Algorithm*", In Proceedings of the IEEE, vol.61, no.3, pp.268-278 1973

[4] Young, S., Evermann, G., Odell, J., Ollason, D., Woodland, P., Kershaw, D., Moore, G., Valtchev, V. „*The HTK Book (for HTK Version 3.1)*", Cambridge University 2001

Prosody Model and its Application to Czech TTS System

Jan Romportl, Jindřich Matoušek, Daniel Tihelka

University of West Bohemia in Pilsen, Department of Cybernetics,
Univerzitní 22, 306 14 Plzeň, Czech Republic
rompi@students.zcu.cz, jmatouse@kky.zcu.cz, dtihelka@kky.zcu.cz

ABSTRACT

The first part of this paper[1] proposes a formal theoretical framework for prosody description. This framework is based on empirically acquired axioms (following the linguistic structuralism) and in terms of the mathematical set theory it presents prosody as a relation between abstract sentence underlaying structures and intonation (together with timing). On the basis of this framework one can utilise various system description and analysis methods as well as pattern processing techniques to model the aforementioned relation. The second part of this text introduces the application of this framework to the Czech text-to-speech system ARTIC. It uses rules to place abstract intonation schemes (melodemes and cadences) depending on the position of an intonation centre of an utterance.

1. INTRODUCTION

Probably all text-to-speech (TTS) concerned papers agree that naturalness and also intelligibility of synthetic speech strongly depends on its prosodic quality. However, if one asks what such "prosodic quality" means one usually gets an answer vaguely summarising the goal of all prosody research to make TTS sound "human-like", no matter methods involved and long-term perspective offered. Indeed, this is under certain circumstances true but we feel that more comprehensive insight to this problematics is needed if the aforementioned dream of all TTS designers should be fulfilled. In this paper we present results of the initial stage of our research on prosody which tries to employ some results achieved by the functional approach of Prague Linguistic School. Couple of our results were utilised and applied to the state-of-art Czech TTS system being developed at Cybernetics department of the University of West Bohemia in Pilsen.

2. PROSODY PROPERTIES

Although "prosody" is generally known as a sort of synonym for suprasegmental features of human speech, we will try to give more formal conception. First of all we must cope with some empirical observations as well as set up what we want to achieve. The latter is explicated by an assumption, that "acceptably natural prosody" means such suprasegmental properties of synthetic speech that would be produced by a human speaker in a clear and intelligible utterance without excessive emotional concern. The former can be briefly summarised by following:

Axiom 1 (based on [1]):

I. Every continuous speech is divisible into smaller units (we will call them "phonemic clauses").
II. These units have their own specific intonation (e.g. melody and intensity, or contour of fundamental frequency and volume) and timing. The word "intonation" is used for the attribute constituted by melody and intensity.
III. These units can be separated by pauses.

Axiom 2:

"Prosodic quality" of every utterance can be fully described by intonation and timing.

Axiom 3 (for discussion see [1], [2], [3]):

Intonation and timing are "functionally involved"; they are constituted by elements where some of them have linguistic function(s) meanwhile some of them do not. Most relevant functions are delimitative and semantical.

The semantical function is (at least) of two kinds: it helps a listener create a notion of an utterance's *meaning* (as a linguistic concept) and participates in creating an utterance's (ontological) *content* (or factual knowledge) which can poorly be derived from the text itself.

Axiom 4 (see [1]):

The principle of tolerance and relevance. Each element of a certain functional layer can realise itself freely within boundaries given by a system.

We do not have enough space to discuss the above axioms and we are aware they might be somehow modified when facing future research. Further in the text we will use this notation: $\langle x_1, x_2, \ldots x_n \rangle$ is an ordered n-tuple of objects $x_1 \ldots x_n$ (in this order); relation R is a set of all 2-tuples $\langle y, x \rangle$ such that objects y and x (in this order) are in a relation R (e.g. yRx); for a relation R symbol *dom*(R) is a set of all x such that $\langle y, x \rangle \in R$; for a relation R symbol *rng*(R) is a set of all y such that $\langle y, x \rangle \in R$; for a set A symbol *pot*(A) is a set of all

1 This research is supported by the Grant Agency of Czech Republic no. 102/02/0124 and the Ministry of Education of Czech Republic, project no. MSM235200004.

subsets of A; for sets A and B operator A | B produces a set of <y, x> such that $\langle y,x\rangle \in A|B \Leftrightarrow \langle y,x\rangle \in A \wedge y \in B$ (notation is based on [10] and slightly modified).

Now we can propose the following definition:

Definition 1:

Be relation P', $<IT_{SJ}, <TR_S, MR_S, A>> \in P'$, called *prosody*. TR_S is the tectogrammatical representation of a sentence S, MR_S is the morphonological representation of a sentence S, A stands for stochastic attributes (perhaps properties that cannot be modeled otherwise), IT_{SJ} is the set whose elements adequately describe intonation and timing of sentence S realised as an uttered sentence token J.

Tectogrammatical representation is such a (non-linear) representation of a sentence which describes both its meaning and its syntax in terms of so-called *tectogrammatical layer* of language description. In detail it is elaborated and described in [4], [5]. It can be also called a *semantical structure*. Morphonological form is understood as a form which represents a sentence as it appears in its very surface structure (more specially for our purposes we can call this a graphemical form). "Adequate description of intonation and timing" is a very vague term but this way we leave the question of suitability of various techniques of intonation and timing description open.

The "real" nature of the aforementioned relation is somehow *infinite* (note the original meaning of this word is closely related to *indeterminate*) thus it would be of great benefit to utilise some mathematical theory for indeterminacy description (see very promising [10]). So far we must settle for the following (maybe contra-intuitive) assumption which helps us establish some sort of "discourse universe":

Axiom 5:

For each 3-tuple $<TR_S, MR_S, A>$ all possible uttered sentence tokens have been brought to existence.

Theorem 1:

$\forall S \in dom(P') \exists IT_{SI}, IT_{SJ} \in rng(P'):$

$\langle S, IT_{SI}\rangle \in P' \wedge \langle S, IT_{SJ}\rangle \in P' \Rightarrow IT_{SI} \neq IT_{SJ}$

Proof is a direct consequence of axioms 4 and 5. In short it formalises the fact that one sentence can be uttered in more ways (concerning intonation and timing).

Definition 2:

Acceptably natural prosody is the relation P = P' | Q where

$Q \in pot(rng(P'))$ such that

$\forall S \in dom(P') \forall q \in rng(P'):$

$q \in Q \Leftrightarrow q = \underset{q_i:\langle S,q_i\rangle \in P'}{argmin}\ J(S, q_i)$

where $J(S, q_i)$ is a criterial function such that (in case of suitable indexing) $J(S, q_1) < J(S, q_2) < ... < J(S, q_n)$.

Theorem 2:

P is a function.

Proof results from the definition 2 and from the requirements posed on the criterial function $J(S, q_i)$ because for each $S \in dom(P)$ exactly one IT_S is given.

The responsibility for constituting functional (not in terms of linguistics but in terms of mathematics) dependency between prosody and text is thus put on the criterial function that can be represented for example by some subjective perceptional tests. Basically its goal is to choose one realization of intonation and timing that is best in terms of given criteria. Technically for the sake of artificial prosody generation (in TTS systems for instance) we may outline the criterial function so as to select such realizations which are as close to specific real data as possible.

Through the above approach we postulated and formalised functional dependency between intonation (plus timing) and abstract linguistic representation of a sentence which can be derived from graphemical (i.e. generally morphonological) representation of this sentence and its context. The reason we considered the tectogrammatical level of representation as underlying instead of some "more surface" level is straightforward: the tectogrammatical representation (TR) is able to describe contextual information with regard to the relevance of *topic-focus articulation* (TFA) which reflects communicative function of a sentence (this can deal for instance with communicational aspects of word ordering which has a significant relevance in Slavic languages). For details see [4], [5], [6], [8]. We are convinced that ignoring all these aspects of text could bring some short-term advantages but seems to be shortsighted when facing the long-term goal of cognitive sciences.

Our framework allows us to understand prosody in terms of the system theory and thus model it using methods based on system description and analysis (in practical applications a lot of work can be done by pattern processing techniques). For the importance of language structures formalising see for example [7].

3. APPLICATION FOR CZECH TTS

Concerning the above mentioned we distinguish two levels of prosody description (see [1], [2], [3], following the Prague linguistic structuralism): functionally motivated and acoustically motivated (though this term might not be best fitting). We are far from thinking some great progress can be achieved without co-operation of these two constituents so we adopted and slightly changed terminology as it is presented in [3].

Continuous text - in written form - is segmented into sentences which we understand to be particular utterances (to make the problem easier) - in spoken form. Each utterance is divided into major and (optionally) minor phonemic clauses (also called prosodic phrases) which are then rhythmically segmented into phonemic words (a phonemic word is one or more words subordinated to one word-stress). We can actually say it

is almost a rule in Czech to place a word stress in the beginning of a phonemic word and for the sake of TTS we can postulate it (in this case we must cope with sometimes occurring "pre-phonemic words"). An example:

'V posledních **'dnech** /$_1$ **'naší 'do**volené /$_2$ jsem **'ko**nečně **'po**chopil //$_2$ že **'ne**snáším **'ces**tování..
In the last days /$_1$ of our holiday /$_2$ I finally realised //$_2$ that I hate traveling.

Major phonemic clauses are divided by /$_2$, minor phonemic clauses by /$_1$ (these are quite optional and strongly depends on speech rate) and // signs pauses. Bold characters show stressed syllables and ' stands for phonemic word beginnings. So far we have implemented rather simple rule-based method of phonemic words and word stresses placement but it has proved to be efficient enough because of relative simplicity of this area in Czech language (word stress has only a delimitative function and does not change the meaning anyway). Phonemic clauses detection is being developed and has not been implemented sufficiently yet.

Further in this text we will speak mostly about intonation (and more specifically about melody) since for Czech it is the most important attribute of prosody (see [1], [3]) and the only one implemented in the TTS system ARTIC so far (see [9]). There is also no tectogrammatical parser available for us which means that at this stage of research and development we cannot fulfil the goals appointed by the theoretical part of this paper. However, they (together with the above formalization) should be kept in mind as well as the fact we want to develop a prosody model for TTS, not describe and formalise language system.

Prosody of an utterance is described (as it was already mentioned) in two levels. The first level consists of so called *melodemes*. Melodeme is an abstract intonational pattern established in certain function within a language system. The second level is characterised purely by acoustic attributes irrespective to their function. If we stay under "protective wings" of *acceptably natural prosody* we can describe this level in terms of *cadences*. We can understand cadence to be a model generating elements of IT_{SJ} set, but - an important note - the input of such model must be a structure much more simple than <TR_S, MR_S, A>. In short - cadence describes "real" acoustic properties of intonation without regard to a meaning or content of an utterance.

So far we use simple rule-based model of cadences of this inventory (in terms of the F0 contour slope): flat, ascending, descending, ascending-descending, descending-ascending. Each cadence can be enhanced by the attribute "stressed" (e.g. it models F0 contour with a word stress) and is presented (except for the flat cadence) in three forms (simply distinguished by numbers 1, 2, 3) which vaguely express "how quick F0 moves up or down" (we will refer to it as the tendency of a cadence). Each cadence places a number of F0 values (in Hertz) at certain timestamps and all such values are then linearly interpolated producing the final F0 contour. These cadences are intended to underlay phonemic words. It means that a cadence is quite a simple unit covering only one phonemic word and thus concerning just morphonological form of a phonemic word (e.g. a subset of MR_S). For example the flat stressed cadence places three values: first at the beginning of the first phoneme of a phonemic word, then local maximum at the beginning of the nucleus of the stressed syllable, then local minimum at the end of the last phoneme of a phonemic word. The concrete numbers are calculated using referential F0 value (supplied by the superordinate layer - melodemes, as will be seen later) and coefficients specific to particular cadence (indeed the coefficients are dependent on a speaker and are set up experimentally).

Melodeme placement depends on the position of so called *intonation centre*. Normal position of an intonation centre in Czech is on the last phonemic word of an utterance. This position is automatically understood to designate emotionally neutral utterance. However, contextual factors of TFA (especially word order) may change the position and this can be expressed by TR_S (for details see [4], [6], [8]). Since we do not have a tectogrammatical parser so far (as it was already mentioned) we must actually omit the structure given by TR_S and we place intonation centre always on its "normal" position. There are also phonemic clause intonation centers and these are automatically placed on the last phonemic word of phonemic clauses.

Intonation centre - because of its function - is a border between two neighbouring melodemes (indeed it is just a conceptual point of view - the function and the intonation centre itself is actually realised by particular melodemes or cadences respectively). This means that an intonation centre starts a new melodeme of the type dependent on a sentence modality (declarative, interrogative, imperative, etc.) and other aspects (not realised in this phase yet). We currently use following melodemes (based on [2], [3]):

M0 - null melodeme
M1 - terminating descending melodeme
  M1-1 - neutral
M2 - terminating ascending melodeme
  M2-1 - neutral
M3 - nonterminating melodeme
  M3-1 - neutral
  M3-2 - neutral, pause preceding

Indeed we should distinguish more melodemes but those written above are quite sufficient for emotionally neutral utterance description (when almost omitting TR_S, and accounting only for a sentence modality as it is our case at this stage of research and prosody application). M0 is understood as a formal description of such a segment of a phonemic clause where actually no melodeme is realised. It ranges from the beginning of a phonemic clause to the phonemic word right before the intonation centre of the phonemic clause and it describes slightly descending tendency of phonemic clause melody. M1 is used for an utterance terminating in case the sentence is declarative, imperative or interrogative-

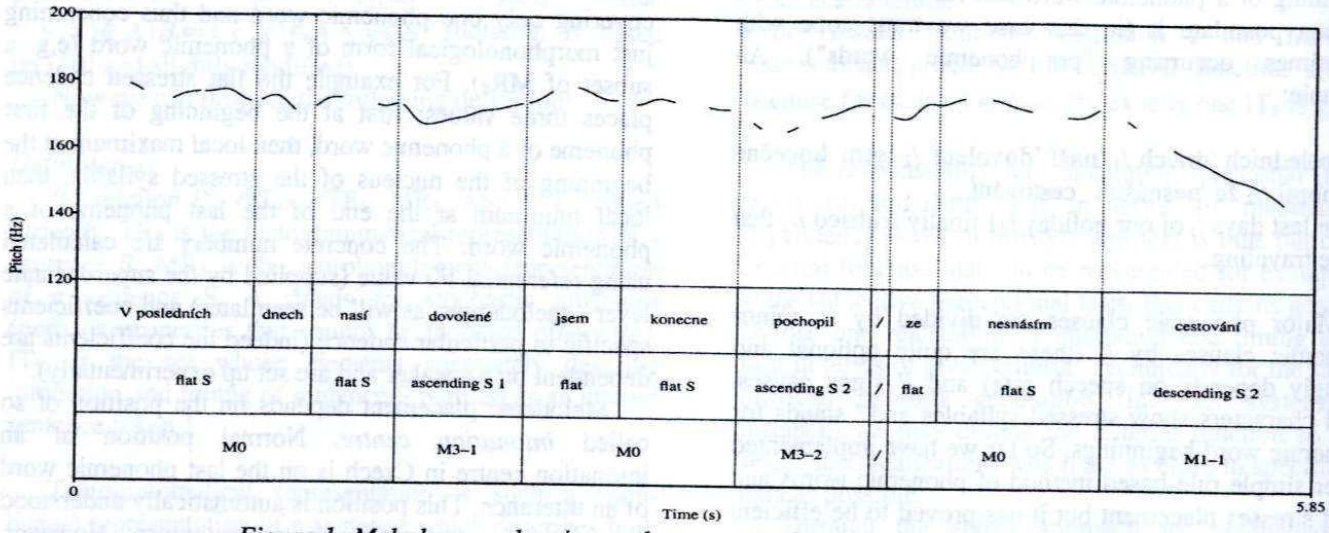

Figure 1: Melodeme and cadence placement in the resulting synthesised speech

supplementary (like English "wh-" questions). It starts at the intonation centre and reaches the end of the utterance. M2 is very similar to M1 with the difference that it is used to terminate interrogative-inquiring sentence (e.g. questions with a "yes/no" answer). M3 is used in those phonemic clauses which are not last in particular utterances.

Melodemes are chosen to fulfil particular function given by the communicative intention and cadences can be understood as tools intended for their concrete realization. Figure 1 transparently illustrates the usage of cadences and melodemes on the example of the above introduced sentence (without minor phonemic clauses).

4. CONCLUSION

The formalization of prosody in the part 2 should not be taken linguistically rigorous. It is certainly purpose build, however, it tries to establish solid ground on which one can base prosody models applied in many systems such as TTS. Moreover, it implies the dependency between the semantical structure and a context of a sentence which is necessary to take into account to build as much as possible human-like sounding TTS system.

The presented rule-based method of prosody description is a very simplified realization of this formalization. However, the results - as can be evaluated with Czech TTS ARTIC system - are quite satisfactory, taking the simplicity of so far used method into account. The intonation is quite far from being perfectly human-like but this can be much improved by the use of more sophisticated cadence models (for example stochastic or neural network based). Moreover, since the prosody description is separated into two autonomous parts (functional and non-functional), changes made to one part almost do not influence the other one and the formalization sets up a solid area for a further research.

The future work will focus on more elaborate cadence models (based on the stochastic system description) and will enhance timing properties, especially syllable duration. More thorough insight into the intensity attribute of intonation will be undertaken too (this means modelling intensity contour not only at stressed syllables).

5. REFERENCES

1. Daneš, F. "Intonace a věta ve spisovné češtině (Sentence Intonation in Present-Day Standard Czech)" , Nakladatelství Československé akademie věd, Prague, 1957
2. Romportl, M. "Základy fonetiky (Basics of Phonetics)", Prague, 1973
3. Palková, Z. "Fonetika a fonologie češtiny (Phonetics and Phonology of Czech)", Karolinum, Prague, 1994
4. Sgall, P. - Hajičová, E. - Panevová, J. "The Meaning of the Sentence in Its Semantic and Pragmatic Aspects", Reider, Dordrecht, 1986
5. Panevová, J. "Formy a funkce ve stavbě české věty (Forms and Function in Czech Syntax)", Academia, Prague, 1980
6. Sgall, P. - Hajičová, E. - Buráňová, E. "Aktuální členění věty v češtině (Topic and Focus in Czech)", Academia, Prague, 1980
7. Hajič, J. - Hajičová, E. - Rosen, A. "Formal Representation of Language Structures", In TELRI Newsletter No.3, pp. 12-19, June 1996
8. Hajičová, E. "The Prague Dependency Treebank: Crossing the Sentence Boundary", In Proceedings of the Second Workshop on Text, Speech, Dialogue, pp. 20-27, Marianske Lazne, Czech Republic, 1999
9. Matoušek, J. - Psutka, J. "ARTIC: A New Czech Text-to-Speech System Using Statistical Approach to Speech Segment Database Construction", In The Proceedings of the 6th International Conference on Spoken Language Processing ICSLP2000, vol. IV. Beijing, China, 2000, pp. 612-615.
10. Vopěnka, P. "Úvod do matematiky v alternatívnej teórii množín (Introduction to Mathematics in Alternative Set Theory)", Alfa, Bratislava, 1989

АНАЛІЗ ТЕМПОРАЛЬНИХ ПЕРЕТВОРЕНЬ МОВНИХ ЕЛЕМЕНТІВ ДЛЯ ЗАДАЧ ЧАСОВОГО МАСШТАБУВАННЯ ГОЛОСОВИХ ПОВІДОМЛЕНЬ

Зореслава Шпак, Юрій Рашкевич

Національний університет "Львівська політехніка"
вул. Ст. Бандери, 12, м. Львів-13, 79646; тел.: (38-0322) 398-793; факс: 744-143
Електронна пошта: zshpak@polynet.lviv.ua; rashkev@polynet.lviv.ua

ABSTRACT

The characteristics of the speech units (sounds, syllables, words, etc.) duration have been specified for different pronunciation types. Time-scale transformation regularity of the Ukrainian voice phonemes sounding at different speech rates has been investigated. Six groups of oral speech elements are determined: stressed vowels, unstressed vowels, voiced consonants, unvoiced consonants, plosive stops and word-spacing pauses. The elements of specified groups are characterized by similar temporal transformations. Non-linear analytic function presenting the relative temporal changes of sound duration have been plotted for every group.

1. ВСТУП

Підвищення ефективності усномовної комунікації в людино-машинних системах ставить задачу керування темпом надходження голосової інформації. Щоб забезпечити розбірливість і натуральність звучання мови, відтвореної зі зміненою швидкістю, у процесах часового масштабування голосових записів треба враховувати природні темпоральні властивості мовних елементів [1].

У роботі наведено часові параметри структурних одиниць мовного потоку для різних темпів мовлення. Проаналізовано закономірності темпоральних змін звуків української мови і виділено класи мовних елементів, що характеризуються спільністю перетворень, викликаних зміною швидкості мовлення. Побудовано графіки, що відображають залежність відносної зміни тривалості елементів виділених класів від загального коефіцієнта зміни темпу.

2. ЧАСОВІ ХАРАКТЕРИСТИКИ РІЗНИХ ТЕМПІВ МОВЛЕННЯ

Для визначення темпоральних параметрів мовлення використано набір текстів, які відображають сучасні тенденції у структурі української розмовної, ділової та наукової мови. Кожен з текстів (загальний обсяг мовного матеріалу – 167 речень та відповідно: 2138 слів, 5331 складів і 12315 фонем) промовлявся групою з семи мовців жіночої та чоловічої статі в чотирьох темпах: швидкому, звичайному, повільному та протяжному. У табл.1 наведено темпоральні дані двох мовців (М1 і М2), які відрізняються стилями мовлення та усереднені значення для всієї групи мовців.

Основні підсумки отриманих результатів:

- найбільш стабільною і незалежною від тексту характеристикою швидкості мовлення є параметр "складів за секунду" – його доцільно застосовувати як базову темпоральну характеристику;

Таблиця 1

Часові характеристики основних стилів мовлення

Параметр		Темп мовлення								
		швидкий			звичайний			повільний		
		М1	М2	сер.	М1	М2	сер.	М1	М2	сер.
Середня тривалість	речення (с)	8,16	5,29	5,79	12,43	5,10	7,57	20,34	9,69	12,08
	синтагми (с)	2,37	1,47	1,68	3,53	2,34	2,47	5,64	2,44	3,45
	слова (мс)	552	341	415	734	441	540	1106	556	667
	складу (мс)	221	131	159	302	173	206	419	213	267
	звуку (мс)	94	56	68	129	74	88	179	91	114
Швидкість мовлення	слів/хвилину	96,5	135,8	130,1	64,6	119,7	89,1	39,2	86,1	62,3
	складів/секунду	4,1	6,7	5,3	2,6	4,7	3,7	1,6	3,4	2,5
	звуків/секунду	9,2	16,3	13,2	6,3	11,5	9,0	3,7	8,2	6,2
Коефіцієнт прискорення/сповільнення		1,51	1,43	1,46	1,0	1,0	1,0	1,64	1,39	1,43

• варіативність темпоральних показників мовців є достатньо широкою, особливо у випадках швидкої та повільної вимови. Так, М1 володіє достатньо широким діапазоном зміни темпу і помірною швидкістю мовлення, а М2 може бути віднесений до групи швидких мовців – властивий для нього звичайний темп є близьким до швидкого мовлення М1;

• незважаючи на відмінності в абсолютних параметрах швидкості мовлення, всім мовцям властиві достатньо близькі значення відносної зміни темпу. Середні по всій групі коефіцієнти максимального прискорення та сповільнення складали відповідно 1,46 та 1,48 (коефіцієнти сповільнення характеризувались значно ширшим розкидом значень, ніж коефіцієнти прискорення);

• часові параметри мовця не залежать від його статі, а визначаються властивим даній людині стилем мовлення. Усім мовцям, які не мали спеціальної дикторської практики, важко було довготривало відхилятись від звичайної швидкості вимови – чим довшим був текст, тим меншим ставав загальний коефіцієнт зміни темпу.

3. ТЕМПОРАЛЬНІ ПЕРЕТВОРЕННЯ ПАУЗ І ЗВУКІВ

3.1. Темпоральні перетворення пауз.

Зміна темпу мовлення пов'язана з перерозподілом структури мовного потоку (рис.1). Перехід від звичайної мови до швидкої чи повільної насамперед відображається в зміні кількості та тривалості пауз: кількість пауз зменшується/зростає усереднено в 1,5 разів, а їх сумарна тривалість змінюється відповідно приблизно в 2,3 рази (табл.2). Водночас загальна тривалість звукової частини мовного потоку скорочується/збільшується тільки в 1,3 рази. Найбільший відсоток пауз у повільній мові – біля 30%. Відзначено, що при подальшому сповільненні темпу дещо зростає кількість пауз, але їх середня тривалість майже не змінюється. Як результат – у протяжній мові частка пауз зменшується до 20%.

Довгі синтагматичні паузи загалом характеризуються меншими і стабільнішими темпоральними змінами, ніж короткі міжслівні, які вирізняються найвищим коефіцієнтом варіації та є найбільш

Рис.1. Відсоток пауз і звуків у загальній тривалості мови

залежними від просодичних особливостей мовців. Середнє значення тривалості звукової ділянки між паузами, яке для експериментальних даних було близьким до 0,96 с, зазнає найменших темпоральних змін . Тільки в протяжному, близькому до співочого мовленні цей параметр зростає в 1,5 рази.

3.2. Зміни голосних звуків.

Для аналізу темпоральних перетворень звуків виконано аудіо-візуальний поділ мовних сигналів кожного із записів. Границі звуків контролювались та уточнювались за динамічними спектрограмами, що дозволило виділити складові частини звуків: початкові та кінцеві переходи і серединні квазістаціонарні ділянки.

Основна різниця відзначена в темпоральній поведінці голосних і приголосних (передусім шумних) звуків. Сповільнення темпу завжди проявлялось у зростанні частки голосних звуків – співвідношення голосні:приголосні для швидкого темпу складало 0,42:0,58; для звичайного – 0,45:0,55; для повільного – 0,51:0,49; а для протяжного – 0,59:0,41. У випадках переходу до швидкого чи повільного мовлення найбільше змінювались тривалості наголошених голосних (табл. 3), але в протяжній вимові спостерігалось невелике зменшення інтонаційного співвідношення між наголошеними і ненаголошеними звуками (рис.2). Місце голосного звуку в слові та

Таблиця 2

Темпоральні характеристики роздільчих пауз

Характеристика	Темп мовлення			
	швидкий	звичайний	повільний	протяжний
Тривалість пауз у 1с мовного потоку (мс)	95	198	293	214
Кількість пауз на слово	0,41	0,56	0,81	0,97
Середня тривалість паузи (мс)	141,6	238,1	383,4	405,7
СКВ тривалості пауз (мс)	129,1	188,7	215,3	273,7
Середня тривалість звукової ділянки між паузами (мс)	1012	664	905	1420
Відносне збільшення загальної тривалості пауз	1,0	2,47	5,73	6,41
Відносне збільшення середньої тривалості паузи	1,0	1,67	2,72	2,84

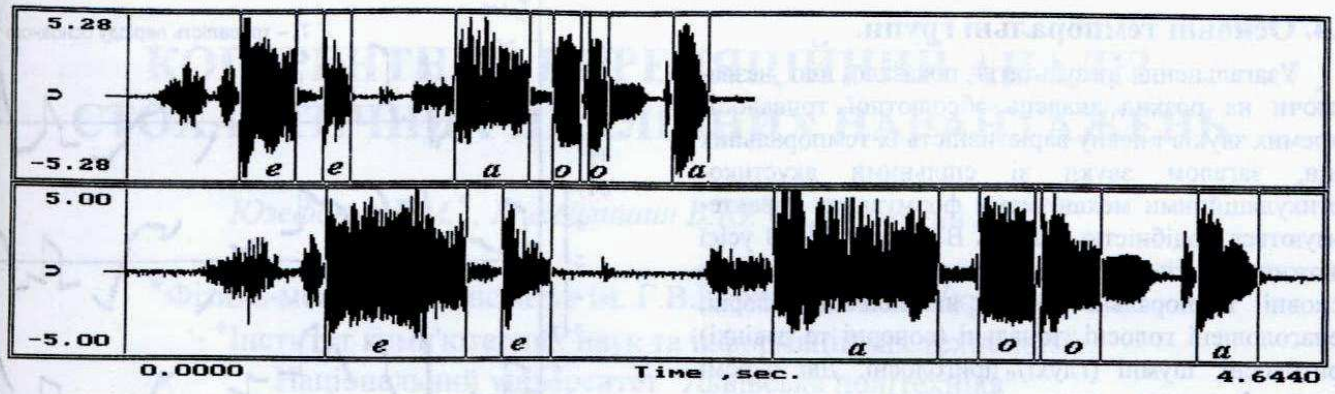

Рис.2. Мовні сигнали словосполучення *щебет жайворонка*, вимовленого в звичайному та протяжному темпах

просодика цього слова мали певний вплив на абсолютну тривалість звуку, проте відносні зміни тривалостей (ВЗТ) всіх голосних, викликані зміною темпу мовлення, були загалом достатньо однорідними.

3.3. Зміни приголосних звуків.

Відмінність у темпоральних перетвореннях голосних і приголосних звуків набуває принципового характеру у випадках значного сповільнення темпу (рис.2). Так, перехід від швидкого до протяжного мовлення пов'язаний із загальним збільшенням тривалості голосних у 4,4 рази, а приголосних – тільки в 2,2 рази. Водночас темпоральна поведінка різних приголосних звуків є теж достатньо неоднорідною (табл. 3). Проведено аналіз темпоральних змін для кожного звуку зокрема та узагальнено для основних класифікаційних груп приголосних [2]. В межах кожної групи враховувався вплив факторів пом'якшення, дзвінкості, назалізації та інших. Зроблено такі основні висновки:

- у випадках невеликих змін темпу (коефіцієнт прискорення/сповільнення не перевищує 1,4) всі приголосні змінюють свою тривалість приблизно однаково; при більших змінах темпу проявляються відмінності в перетвореннях приголосних різної артикуляційної природи;
- перетворення сонорних звуків при невеликих змінах темпу є близькими до перетворень ненаголошених голосних, а при значних змінах темпу вони є ближчими до перетворень дзвінких приголосних;
- темпоральні зміни зімкнених звуків проявляються в неоднаковій зміні їх складових: паузи зімкнення та наступної короткої звукової ділянки (табл. 3);
- як для сонорних, так і для щілинних та зімкнених звуків не встановлено істотних відмінностей у темпоральних змінах тривалостей м'яких та відповідних твердих звуків, за винятком випадків, коли м'які приголосні виступали як подовжені;
- приголосні з тональною складовою (дзвінкі звуки) дещо відрізняються за параметром тривалості, а також за значеннями ВЗТ від відповідних глухих приголосних (табл. 3), особливо у випадках протяжного мовлення;
- збільшення тривалості початково довгих звуків (зокрема подовжених приголосних) і скорочення початково коротких (зокрема вибухових) відбувається в меншій мірі, ніж для інших звуків даної групи.

Таблиця 3

Характеристики темпоральної тривалості звуків базових класів

Класи звуків				Темп мовлення										
				швидкий			звичайний			повільний			протяжний		
				ПВ, %	СТ, мс	ВЗТ	ПВ, %	СТ, мс	ВЗТ	ПВ, %	СТ, мс	ВЗТ	ПВ, %	СТ, мс	ВЗТ
Голосні	ненаголошені			22,6	59,5	1,32	22,8	78,8	1,0	25,0	105,8	1,34	29,6	239,5	3,04
	наголошені			19,5	91,1	1,42	21,9	129,6	1,0	25,2	181,5	1,40	29,8	408,5	3,15
Приголосні	сонорні			20,6	49,6	1,30	19,9	64,7	1,0	17,5	84,6	1,31	15,1	119,9	1,85
	шумні	щілинні	дзвінкі	4,8	61,7	1,27	4,6	78,6	1,0	4,1	98,9	1,26	3,4	136,8	1,74
			глухі	8,9	99,4	1,26	8,7	125,6	1,0	7,7	146,5	1,17	5,3	168,1	1,34
		зімкнені	дзвінкі	4,8	36,5	1,32	4,7	48,0	1,0	4,0	60,9	1,27	3,2	79,5	1,66
			глухі	2,6	24,6	1,25	2,5	30,7	1,0	2,1	36,6	1,19	1,2	42,3	1,38
			змикання	10,2	71,0	1,27	10,6	90,1	1,0	10,8	109,1	1,21	8,3	146,4	1,62
		африкати		2,8	77,1	1,28	2,6	98,4	1,0	2,4	110,8	1,13	1,7	127,6	1,30
	подовжені			1,7	108,7	1,31	1,7	142,4	1,0	1,6	178,3	1,25	1,3	216,8	1,52

Примітка: ПВ - процентний вміст; СТ - середня тривалість; ВЗТ - відносна зміна (збільшення/зменшення) тривалості.

3.4. Основні темпоральні групи.

Узагальнення результатів показало, що незважаючи на розкид значень абсолютної тривалості окремих звуків і певну варіативність їх темпоральних змін, загалом звуки зі спільними акустико-артикуляційними механізмами формування характеризуються подібністю значень ВЗТ (табл. 3). З усієї множини звуків української мови виділено чотири основні темпоральні групи: наголошені голосні, ненаголошені голосні, тональні (сонорні та дзвінкі) приголосні, шумні (глухі) приголосні. Дві окремі групи складають паузи зімкнення та роздільчі (міжслівні) паузи, темпоральні модифікації яких є принципово різними. На рис.3 наведено графіки ВЗТ мовних елементів основних темпоральних груп відносно значень, властивих для швидкого темпу (k – загальний коефіцієнт сповільнення вимови).

Зміна швидкості мовлення викликає неоднакові перетворення стаціонарної і перехідних ділянок звуків. Якщо зміни тривалості стаціонарів є достатньо регулярними, то трансформації перехідних ділянок характеризуються великою варіативністю. Загалом змінюються тривалості обох переходів, але усереднено в 1,4 (а в разі значних сповільнень – в 1,7) рази менше, ніж стаціонарних частин звуків.

В процесі досліджень проаналізовано вплив темпу мовлення на значення тривалості періодів основного тону (ПОТ) звуків, а також на форму мелодичного контуру фраз. Встановлено, що середні значення тривалості ПОТ тональних звуків і мелодика мови є практично незалежними від швидкості мовлення (рис.4). Зафіксовано тільки невелике зростання (біля 4%) середньої тривалості ПОТ у протяжному темпі та модуляції значень ПОТ для деяких довгих голосних. Темпорально незалежними виявилися також розподіли формантних частот і спектральні структури шумних звуків.

4. ВИСНОВКИ

Природна зміна швидкості мовлення пов'язана зі зміною тривалості всіх елементів мовного потоку. У разі невеликих змін темпу основних трансформацій зазнають паузи, а модифікації всіх звуків є достатньо подібними і можуть бути описані лінійною залежністю від коефіцієнта зміни темпу. У випадках сповільнення/прискорення більше, ніж у 1,5 рази проявляються розбіжності у темпоральних перетвореннях звуків з різними механізмами формування. Ці відмінності стають принциповими для протяжної мови.

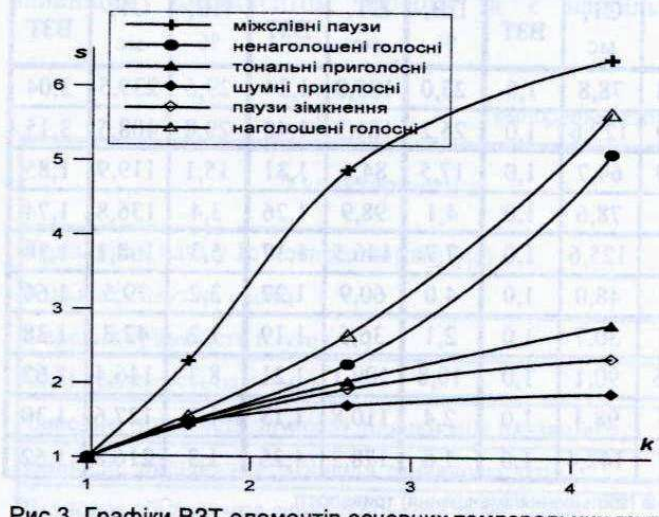

Рис.3. Графіки ВЗТ елементів основних темпоральних груп

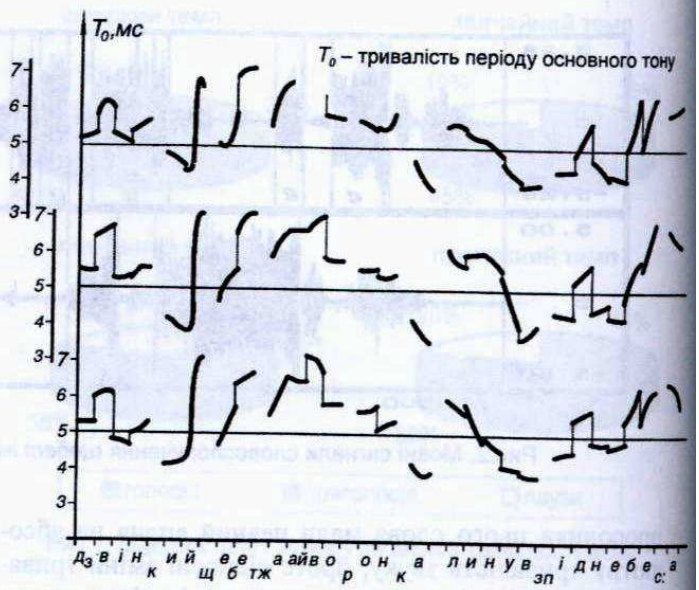

Рис.4. Мелодичні контури тестової фрази у звичайному (нижній рисунок), повільному (середній) і протяжному (верхній) темпах

За близькістю значень ВЗТ і подібністю перетворення структур сигналів, викликаних зміною швидкості мовлення, всі звуки української мови можна об'єднати в чотири основні темпоральні групи; ще дві окремі групи формують паузи. Залежність ВЗТ елементів кожної з цих груп від величини загального коефіцієнта зміни темпу носить нелінійний характер. Особливості зміни тривалості надмірно довгих чи коротких звуків вимагають додаткового введення функції нормування [3]. Підтверджено, що основні спектральні характеристики звуків є темпорально стабільними.

Таким чином, реалізацію зміни швидкості відтворення мовних записів за умови натуральності звучання голосового повідомлення необхідно здійснювати через диференційовану зміну тривалості мовних елементів відповідно до властивих їм значень ВЗТ, зберігаючи загальну структуру звукових сигналів, передусім тональних.

ЛІТЕРАТУРА

1. Рашкевич Ю.М. *Перетворення часового масштабу мовних сигналів.* - Львів: Академічний експрес, 1997. - 140 с.

2. Тоцька Н.І. *Сучасна українська літературна мова.* – Київ: Вища школа, 1981. – 183 с.

3. З. Шпак. *Нормування тривалостей звуків у процесі сповільненого відтворення мовних повідомлень* // Вісник Націон. ун-ту "Львівська політехніка": Львів, 2001. - №433. - С. 266-271.