Intensity Feature for Speech Stress Detection

László Czap
Department of Automation and Info-communication
University of Miskolc
Miskolc, Hungary
czap@uni-miskolc.hu

Judit Mária Pintér
Department of Automation and Info-communication
University of Miskolc
Miskolc, Hungary
pinterjm@uni-miskolc.hu

***Abstract*— Suprasegmental features are fundamental properties of speech. They can improve not only the naturalness of synthesized speech, but the performance of machine speech recognition in voice controlled logistic systems. In linguistics, stress is the relative emphasis that may be given to certain syllables in a word, or to certain words in a phrase or sentence. The term is also used for similar patterns of phonetic prominence inside syllables. The stress placed on syllables within words is called word stress or lexical stress. The stress placed on words within sentences is called sentence stress or prosodic stress. Sentence and word stress are crucial prosodic features. One of the features usually used for stress detection is the energy of syllables, but the average energies of vowels are various. Energy of a stressed weak vowel can be lower than that of an unstressed strong vowel. We compare the amplitude of the actual vowel to that of its average to show the stressed or unstressed nature of the syllable. Average energies of vowels are obtained from a speech recognizer trained with voices of hundreds of speakers.**

Keywords*— stress detection, speech recognition, acoustic feature extraction methods

I. Introduction

The role of stress in human-human communication is various. Acoustic prosodic features are used to make distinction between lexical items creating different stress patterns for words that contain the same phonemes. Prosodic features structure utterances in terms of phrases and to indicate relationships between phrases of sentences. For example, utterances containing lists have their own special prosodic structure. Stress can focus attention on certain words for a variety of purposes, such as highlighting a contrast, emphasizing their importance, or enhancing their intelligibility. All these properties help the listener to understand the contents of a message. In the preverbal childhood, all communication is purely prosodic, and prosody remains important feature in communication.

Traditional statistical continuous speech recognition – such as speech recognition based on hidden Markov models (HMM) – takes the speech as a series of a particular recognition unit (eg. phonemes, diphones, syllables, words, etc.). In statistic based recognition after the process of extracting only the relevant features with reduced redundancy compose the input to the recognition. However, speech is not only a series of these recognition units, but stress, intonation and other suprasegmental speech features are also important. Lexical stress is an important property for the Hungarian language, which can be used to provide lexical constraints or to improve acoustic modeling for automatic speech recognition.

II. Suprasegmental Speech Features

Suprasegmental features have additional information that will contribute to understand the speech. By these features, the speaker can express emotions, syntactic and pragmatic information, etc. [1]. Suprasegmental features are speech rate, pause, rhythm and tonality, volume, tone, intonation and stress.

Using these features, we can consider word stress, section stress (word structures) and sentence stress as units. The assumption that prosodic emphasis is equally likely to fall on any word in an utterance is unjustified, given that primary stress is typically reserved for content words, whereas function words are rarely accented. Word stress means highlighting a syllable of a word, differentiating it from the other syllables of the words.

Considering stress, languages can be classified into two groups:

- bound (stress is always on a clearly identified syllable of the word) ;
- and unbound (variable stress).

Hungarian language is bound, since the stress realizes always on the first syllable (its role is to highlight the essential elements of communication and to articulate logically the communication).

Expressing strong emotions, stress may shift in bound languages, or even can appear on every syllables of a word too. In creation of the stress, three main factors individually or collectively play a role depending on the specific language in which regularities can be observed [2]. These factors are as follows:

- fundamental frequency rises on the stressed syllable (pitch accent);
- higher articulation intensity of the stressed syllable (dynamic accent);
- duration of the vowel of stressed syllable is lengthening (quantitative accent).

Stress may be realized to varying degrees on different words in a sentence; sometimes the difference between the acoustic signals of stressed and unstressed syllables may be

978-1-4799-7370-5/15/$31.00 ©2015 IEEE

minimal. Stressed syllables are often perceived as being more forceful than non-stressed syllables.

III. Stress Detection

A. Stress Detection Methods

The intensity of the detection is usually performed by using the current energy. The values of length, amplitude and spectral changes normalized by using various methods have been analyzed [3]. As regards English language, in many cases automatic stress detection has been implemented by training and using deep neural networks [4, 5]. In development of an English language teaching software that examines the stress patterns produced by the learners, correcting them in given cases. Some of researchers reached the conclusion that combination of length and amplitude is the most reliable sign of stress [5]. As regards the Hungarian language, the stress or indirectly stress determining acoustic-prosodic features (pitch and energy) enable word segmentation [6]. Using a proper feature extraction method, average energy of individual sounds varies greatly, so the energy of low-energy, stressed vowels does not reach that of the high-energy, unstressed vowels.

B. Hungarian Speech Database

We have carried out stress detection tests on phone-based speech recognition trained by Hungarian Speech Database and tested by the first Database for stress distribution in Hungarian sentences. The train database consists of read text, has been compiled in accordance with special phonetic expectations, and has been recorded in average user environment (offices, laboratories, flats), it has been therefore suitable for carrying out targeted investigations.

- Technical data of HSDb:
- • Hungarian, read text, recorded by personal computer;
- • 16 bit, 16 kHz sampling;
- • 300 speakers have recorded the audio files directly by using computer;
- • 50-60 second of speech from a speaker
- • audio samples have been recorded by using different microphones, sound cards and PC-s;
- • variable noisy office, laboratory, home environment;
- the full content of the database has been annotated.

The test database is a reference Hungarian sentence database, where the accent distribution within the sentences is shown by using stress markers in the text of the sentence. An auditive form of every sentence is also available together with the word boundaries within the sentence. The database can be used for research and teaching as well. The database consists of 1866 declarative Hungarian written sentences gathered from literature. The spoken forms of the sentences were

We used 50 sentences from 10 different speakers for testing and analyzing [7].

C. Examination of Different Feature Extraction Methods

In automatic speech recognition, an acoustic preprocessing unit is used in acoustic level that analyses, highlights and compresses the speech signal. The output will be time frame parameters (specific vectors), a discrete vector sequence with a given dimension generated by the digitized speech signal [8].

We represent the energy of a syllable by the energy of the stationary state of its vowel. Context dependent reference vowel energy can be derived from the first state of the Markov model of a diphone, starting with the vowel in the consideration. Regarding the energies, it can be shown that an unstressed high-energy vowel's energy can be higher than that of a stressed low-energy vowel. Therefore, you should compare the energy of the current vowel to the average energy of the vowel of corresponding diphone that can be extracted from the states of the trained HMM.

Procedure of creating relative intensity:

1. Training an HMM recognizer with the HSDb database.
2. Context dependent reference energy of a vowel is derived from the first state of the Markov model of the certain diphone.
3. Current energy of a vowel is got from feature extraction, segmentation to diphones is done by forced alignment and manually corrected.
4. Relative intensity is the ratio of current vowel energy and the reference one.

The efficiency and segmentation accuracy of the trained speech recognizer and the correct calculation of the log energy is strongly dependent on correct timing, that influenced the feature extraction method. On the Table 1 and on the Fig.1 we can see the results of the segmentation error test.

The tested feature extraction methods:

- MFC: Mel-Frequency Cepstral Coefficients (MFCC) is a representation of the real cepstral of a windowed short time signal derived from the Fast Fourier Transform (FFT) of that signal [9];
- PLP: The Perceptual Linear Prediction PLP model developed by Hermansky. PLP models the human speech based on the concept of psychophysics of hearing [9];
- MEL: The 125 Hz – 8 kHz frequency range is divided into 30 mel frequency bands.

The accuracy of segmentation within a given tolerance is shown in Figure 1 in the function of tolerance.

1. Table Accuracy of segmentation of different feature extraction methods

	Feature extraction methods		
Time (s)	***MFC***	***PLP***	***MEL***
<= 0,01	0,31	0,48	0,38
<= 0,02	0,61	0,72	0,60
<= 0,03	0,78	0,84	0,73
<= 0,04	0,86	0,90	0,82
<= 0,05	0,90	0,93	0,88
<= 0,06	0,93	0,94	0,91
<= 0,07	0,95	0,95	0,93
<= 0,08	0,95	0,95	0,93
<= 0,09	0,96	0,95	0,94
<= 0,10	0,96	0,96	0,94

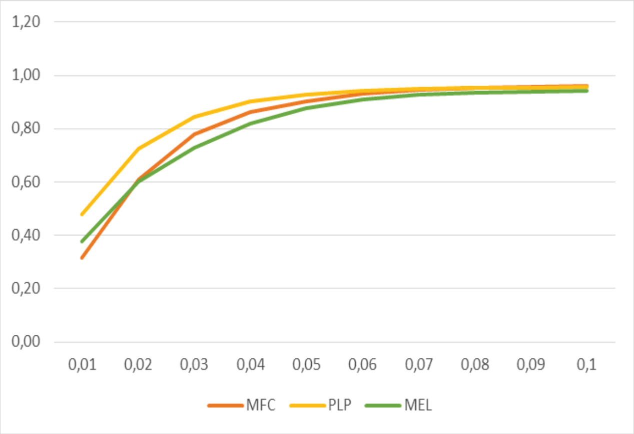

Fig. 1: Toleration of segmentation errors of different features (MFC,PLP,MEL).

We have added the logarithmic energy component to the PLP features (12 dimension PLP, their first and second order derivatives), by which we have been able to examine the overall average energy of each vowel. The first figure illustrates the average energies of vowels extracted from the hidden Markov models of phone based speech recognition trained with the HSDb. These average energies are calculated by averaging the energy of all occurrences of a vowel in the phones.

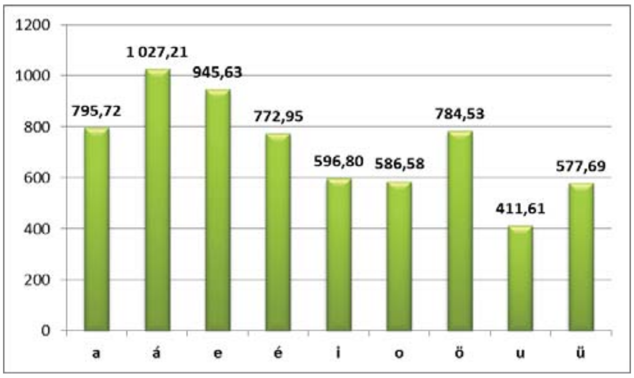

Fig. 2: Average magnitude of vowels in case of PLP feature extraction.

To show the point of our method, we have selected a sentence form the HSDb: 'hod' my:k2dnEk it: ke:rEm O joksObA:jok'. Figure 2. illustrates the momentary Syllable energy of the sample sentence. For reasons of clarity audio files of the example sentence are available at the following link: http://mazsola.iit.uni-miskolc.hu/DATA/research/audio/pelda_mondat.wav.

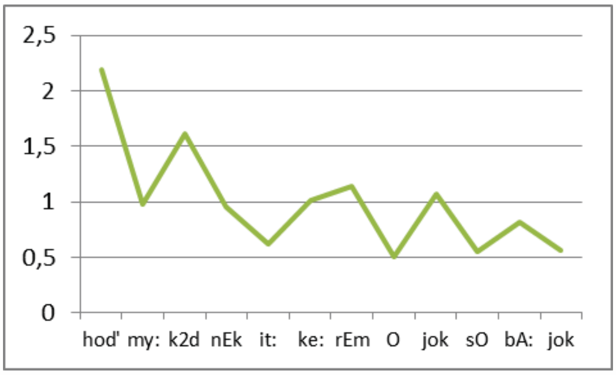

Fig. 3: Measured energy of syllables.

Figure 3. depicts the relative intensity of the same sentence. The context dependent average energy is extracted from the HMM models of dyphones. Current energy levels of the MFCC features of the example sentence have been normalized, that has been suitable for comparison.

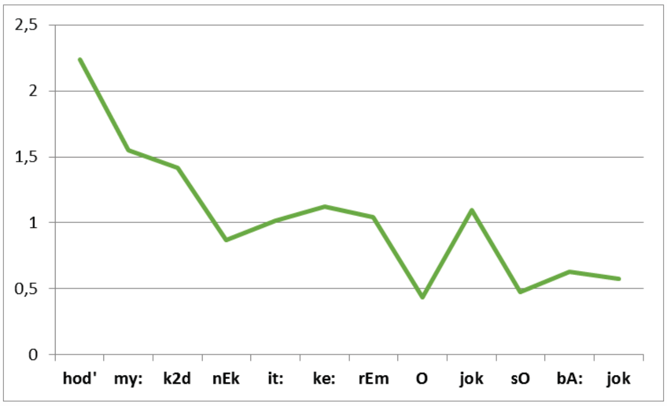

Fig. 4: Relative intensity of syllables.

As it can be detected in case of all sentences, the long term (150 ms) average energy of a sentence is decreasing from the beginning to end. Eliminating this variation of energy, we examined the relative intensity compared to that of its neighbors. Let's call it curve of emergence. To get the emergence curve for the nth syllable:

$$e(n)=(r(n)\div r(n-1)+r(n)\div r(n+1))\div 2 \qquad (1)$$

for the first one:

$$e(1)=r(1)\div r(2) \qquad (2)$$

and for the last one:

$$e(N)=r(N)\div r(N-1) \qquad (3)$$

where r(n) is the relative energy of the nth syllable. Figure 4. shows the curve of emergence for the sample sentence. The primary accent on the first syllable of word joksOba:jok is explicitly highlighted.

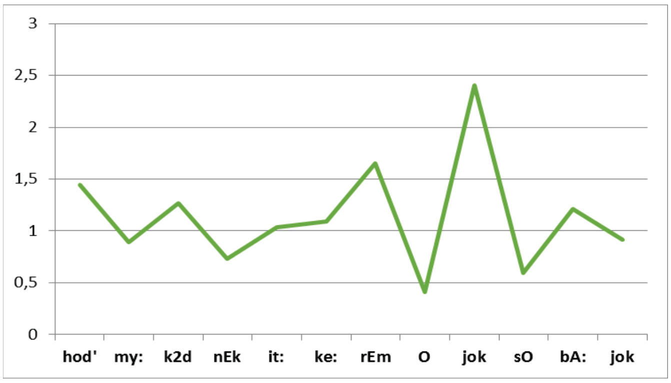

Fig. 5: The curve of emergence for the syllables of sample sentence.

To verify the results, we have compared it to the human judgment of the same sentence. A perceptual subjective test was carried out to detect stressed or unstressed quality of syllables. 16 linguist students of the University of Miskolc were the subjects of the test. When listening to an utterance, students were asked to listen to the acoustic cues but probably could not consider stress as independent of syntactic and/or semantic cues determining the location of stresses. They were ticking off each syllable if they consider it:

- stressed;
- neutral or
- unstressed.

The scores given by subjects were summarized and weighted by:

- 1 for stressed;
- 0.5 for neutral and;
- 0 for unstressed syllables.

These sums are normalized then by dividing them by 16, the number of students. This way we can get stress scores between 1 (definitely stressed) and 0 (explicitly unstressed).

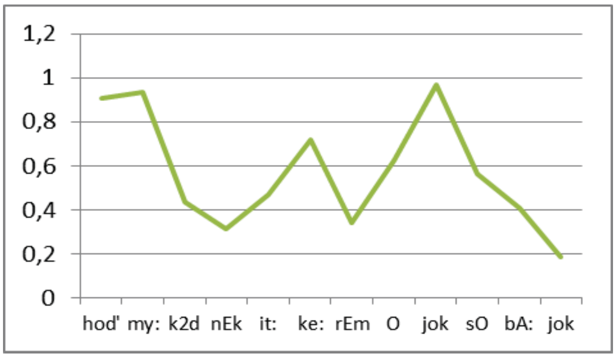

Fig. 6: Stress scores of subjective perception test of sample sentence.

The sentence starts with an interrogative word (hod'). The primary accent is on the first syllable of word joksOba:jok. A secondary accent can be seen on the word ke:rEm.

For comparison purposes we calculated Pearson product-moment correlation coefficient. We have compared from syllable to syllable the momentary energy (Fig. 2.) and the relative intensity (Fig. 3.) to the subjective opinion scores (Fig. 5.)

for momentary energy: re= 0.3811 (4)

for relative intensity: ri= 0.5639 (5)

where re is the Pearson correlation of the momentary energy and the subjective opinion scores, ri is the Pearson correlation of the relative intensity and the subjective opinion scores.

IV. Conclusions

The preliminary results obtained so far show that our method should be further studied to examine intensity component of stress detection. Evaluating the example sentences stressed syllables can be suggested. We are going to examine further feature extraction methods to compare them to the results have been achieved so far, in order to implement the process of automatic evaluation.

Acknowledgment

The research has been carried out within the framework of Mechatronics and Logistics Centre of Excellence operating as one of the strategic research area of the University of Miskolc and as the part of the TÁMOP-4.2.2.C-11/1/KONV-2012-0002 project funded by the European Union, co-financed by the European Social Fund.

References

[1] Gósy M.: Phonetics, the Science of Speech, Osiris, Budapest, 2004. pp.182–243. (In Hungarian.)

[2] Kassai I.: Phonetics, Nemzeti Tankönyvkiadó, Budapest, 1998. (In Hungarian.)

[3] Van Kuijk, D., Boves, L. Acoustic characteristics of lexical stress in continuous telephone speech. Speech Communication, 27(2), 1999, pp.95–112.

[4] Kun L., Xiaojun Q., Shiyin K.., Helen M.: Lexical Stress Detection for L2 English Speech Using Deep Belief Networks, INTERSPEECH, 2013, pp. 1811–1815.

[5] Xie, H., Andrea E, P., Zhang, M., Warren, P.: Detecting stress in spoken English using decision trees and support vector machines. In: Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalization, Australian Computer Society, Inc. 2004, 145–150.

[6] Vicsi, K., Szaszák, GY.: Automatic Segmentation of Continuous Speech on Word Level Based on Supra-segmental Features. International Journal of Speech Technology, Vol. 8, Num. 4, 2005, pp. 363–70.

[7] Abari K, Olaszy G: Magyar hangsúly adatbázis az interneten kutatáshoz és oktatáshoz, X. Hungarian Conference on Computational Linguistics: MSZNY 2014. 396 p.

[8] S. Young.: The HTK Book (for HTK Version 3.3), 2005.

[9] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” Acoustical Society of America Journal, vol. 87, pp.1738–1752, Apr. 1990