From speech corpus to intonation corpus: clustering phrase pitch contours of Lithuanian

Gailius Raškinis, Asta Kazlauskienė

Vytautas Magnus University, K. Donelaičio g. 58, 44248, Kaunas, Lithuania

g.raskinis@if.vdu.lt, a.kazlauskiene@hmf.vdu.lt

ABSTRACT

This paper presents our research in preparation to compile a Lithuanian intonation corpus. The main objective of this research was to discover characteristic patterns of Lithuanian intonation through clustering of pitch contours of intermediate intonation phrases. The paper covers the set of procedures that were used to extend an ordinary speech corpus to make it suitable for intonation analysis. The process of intonation analysis included pitch extraction, pitch normalization, estimation of the representative frequency of a syllable, measurement of an inter-phrase similarity, k-means phrase clustering, and visualisation of clustering results. These computational procedures were applied to 23 hours of read speech containing 41417 phrases. The clustering results revealed some interesting intonation patterns of Lithuanian that could be related to the well known linguistic-prosodic phenomena. Language-independence is an important feature of computational procedures covered by this paper. If speech waveforms and the knowledge of phone and phrase boundaries are given, these procedures can be used for the analysis of intonation of other languages.

KEYWORDS: corpus, prosody, intonation, pitch, syllable, dynamic time warping, k-means clustering, Lithuanian.

1 Introduction

Intonation corpus can be thought of as a speech corpus that has its orthographic / phonetic annotations complemented with prosodic labels. Prosodic labels may characterize both prosodic phenomena themselves (e.g. pitch accents and boundary tones) and features that affect prosody (e.g. logical stress). Some prosodic labels may be associated to a precise timing in speech.

Fuelled by the needs of the text-to-speech synthesis, corpus-based intonation research has been an active research field for more than two decades. Automatic prosodic labelling is among its main topics of interest. Prosodic labels such as word boundary strength (Wightman and Ostendorf, 1994; Vereecken et al., 1998, Heggtveit and Natvig, 2004), stress (Heggtveit and Natvig, 2004), syllable prominence (Wightman and Ostendorf, 1994), word prominence (Vereecken et al., 1998), and pitch accent type (Levow, 2008; Escudero-Mancebo et al, 2012) have been predicted using a variety of supervised machine learning techniques. These techniques included decision trees (Wightman and Ostendorf, 1994), artificial neural nets (Vereecken et al., 1998), classification and regression trees (Heggtveit and Natvig, 2004) sometimes coupled with probabilistic sequential smoothing methods such as Markov models (Wightman and Ostendorf, 1994) and Conditional Random Fields (Levow, 2008).

Supervised machine learning techniques must be trained on a carefully hand-labelled intonation corpora. Manual labelling is usually done by domain experts and represents a tedious and time consuming task. It comes as no surprise that Lithuanian has not any intonation corpus available.

Consequently, corpus-based research of Lithuanian intonation has not received much attention. It has been suggested that Lithuanian is characterized by 7 distinct intonation patterns (Kundrotas, 2008), this suggestion being based on the hypothesis-driven investigative approach inspired by the Lithuanian-Russian intonation parallels (Kundrotas, 2009). Those few audio recordings that have been made during these investigations represent cases illustrating the preliminary set of hypotheses.

Our ultimate research objective is to compile a Lithuanian intonation corpus which could be further explored via supervised machine learning techniques. The main objective of this research is limited in scope and aims to discover characteristic patterns of Lithuanian intonation. In contrast to the abovementioned investigation, we follow a data-driven approach. It assumes that reliable discoveries about intonation patterns of a given language can be made through the computer-aided analysis of a real world intonation phenomena. It would be of a great help to a human expert if similar intonation patterns are collected together and presented to him/her in a user-friendly way. Then, he/she could assign these collections to particular intonation categories, construct “an inventory” of intonation patterns, easily detect and eliminate “false” members out of these collections, mark pitch accents and boundary tones in a more uniform way.

This paper describes computational procedures that allow to find clusters of pitch[1] contours of spoken phrases[2] in a speech corpus consisting of audio records and their orthographic transcriptions. The whole process consists of the following steps:

- manually complementing speech corpus with the word ”break indices” that show the strength of the boundary between two orthographic words.
- automatic alignment of speech waveforms and phonetic symbols (phone level) resulting in the knowledge of both phone boundaries and phrase boundaries.
- estimation of pitch contours of speech waveforms.
- chunking pitch contours to segments corresponding to phrases, and estimating a “representative” pitch for every syllable nucleus of a phrase.
- estimating distances between all possible phrase pairs (on the basis of a pitch).
- grouping phrases by their distances using the k-means clustering technique.
- concatenating clustered data into one audio and one annotation file per cluster.

The steps above are described in more detail in the following sections.

2 Extending speech corpus

We started from the speech corpus that had been used for speech recognition research of Lithuanian (Vaičiūnas, 2006). This corpus consists of audio recordings produced by 50 speakers (25 males, 25 females), each of them speaking for 1 hour. The speech is stored as a collection of 2 min. audio files. The orthographic word-level transcriptions had been prepared to describe the content of audio recordings as accurately as possible.

2.1 Extending speech corpus with word “break indices”

Word “break indices”, representing the strength of the boundary between two orthographic words, were assigned to each pair of contiguous words and were inserted at the “orthographic” level of a text (see fig 1) by a human expert. Our set of indices followed the standard of the ToBI transcription framework (Beckman et. al., 2005).

```
<file name="SIR_A001"> devyni4 kieti4 / plieni4niai skam3balo smū3giai /
paža3dino liu4dą vasa3rį iš_mie3go // krū9ptelėjęs jis_atme9rkė aki4s /
ir_valandė3lę negalė9jo susivo9kti kame4 / <accent correct="e3sąs">esą3s</accent>
// bu4vo ly9giai penkta4 valanda4 ša9lto ruden3s ry9to </file>
```

FIGURE 1: An excerpt of the “extended” orthographic transcription associated to an audio file. Both “ortographic” and “break index” levels of the ToBI framework are mixed together. The symbols / and // indicate an intermediate and a full intonation phrase boundary respectively. The symbol _ indicates the absence of a word boundary (with evidence of cliticisation).

[1] Though the term “pitch“ is a perceptual characteristics, we use it in the sense of fundamental frequency.

[2] We use the short term “phrase” to denote the speech segment within intermediate phrase boundaries.

2.2 Automatic alignment of speech waveforms and phonetic symbols

The process of automatic alignment of speech waveforms and phonetic symbols (phone level) had the objective of identifying timings of phone and phrase boundaries within speech waveforms. The extended orthographic transcription was submitted to the Lithuanian grapheme-to-phoneme converter (Norkevičius et al, 2005) which output the corresponding sequence of phones. Within this sequence, full intonation phrase boundaries were replaced by the "silence" phone, and the optional "short pause" phone was allowed between words.

Given speech waveforms and their corresponding phone sequences the stochastic alignment approach was used. First, speaker-specific (trained on 1 hour of speech) acoustic models were constructed closely following the sequence of processing steps suggested by the HTK book (Young et al., 2000). Thereafter automatic alignment of speech waveforms and phonetic symbols was realized by the Viterbi algorithm which makes part of the same HTK toolkit. Phone and phrase boundaries were identified with the time resolution of 5 ms.

The results of automatic alignment were inspected manually and were found to be satisfactory. Phone boundaries resulting from an automatic alignment procedure are illustrated in fig. 4 (top tier out of three tiers)

3 Intonation analysis and clustering

3.1 Pitch extraction

The following processing steps were used for pitch extraction out of speech waveforms:

- low-pass filtering speech files at cut-off frequency of 1837.5 Hz and downsampling them to 3675 Hz[3].
- estimating the average pitch period *avgp* per speech file on the basis of the downsampled speech in order to define the pitch analysis range of 3 octaves from $1/(4*avgp)$ to $2/avgp$
- low-pass filtering the original speech waveform at a cut-off frequency of $2/avgp + 200$ Hz.
- finding pitch period candidates in this filtered waveform using a cross-correlation technique and time resolution of 2.5 ms (400 Hz sampling rate).
- smoothing pitch contour (selecting one candidate per frame) using the "islands of confidence" approach (Raškinis, 2000)

3.2 Grouping phrases into similarity clusters

3.2.1 Estimating the representative pitch of a syllable

In order to minimize the inter-talker variability (males, females), pitch contours were converted to logarithmic scale (semitones) and zero-centred by subtracting the pitch average of an entire audio file.

[3] The frequency of 3675 Hz was selected for its integer ratios with the common sampling frequencies of 11025Hz, 22050Hz and 44100Hz.

In order to make the intonation description of a phrase more compact the pitch contour of a phrase (sampled at 400 Hz) was replaced by the sequence of syllable pitches of that phrase. The pitch of a syllable varies over time. Thus, every syllable was assigned a "representative" pitch on the basis of the pitch contour over a syllable nucleus (vowel/diphthong). Because of the assumption that the pitch target is best approximated in the later portion of the syllable (Xu, 2004), the representative pitch was calculated by linearly weighting pitch contour values. Pitch contour values at the beginning and the end of a syllable nucleus were de-emphasized and emphasized respectively. Let $f_0[i]$ denote the normalized pitch value of the i^{th} frame, and let $[t_{beg}\ t_{end}]$ be the boundaries of a syllable nucleus, then the frequency representative sf of the syllable was estimated as follows:

$$sf = \frac{\sum_{i=t_{beg}}^{t_{end}} (f_o[i] \cdot \frac{i - t_{beg}}{t_{end} - t_{beg}})}{\sum_{i=t_{beg}}^{t_{end}} \frac{i - t_{beg}}{t_{end} - t_{beg}}}$$

Unvoiced frames (undefined $f_0[i]$ values) were skipped when iterating through pitch contour values. Unvoiced syllables, i.e. syllables for which the pitch detection failed for every frame, were omitted from further consideration.

3.2.2 Defining distance between two phrases

One of the main problems in estimating the similarity of intonation between two phrases was the large variation in their syllable numbers (see fig. 2). Let $F_1 = \{sf_{1,1}, sf_{1,2}, ..., sf_{1,n1}\}$ and $F_2 = \{sf_{2,1}, sf_{2,2}, ..., sf_{2,n2}\}$ be two phrases consisting of sequences of representative syllable pitches, where n_1 and n_2 denote the number of syllables in the phrases F_1 and F_2 respectively. Let $d(i, j)$ denote the distance between truncated sequences $\{sf_{1,1}, sf_{1,2}, ..., sf_{1,i}\}$ and $\{sf_{2,1}, sf_{2,2}, ..., sf_{2,j}\}$ both corresponding to initial segments of F_1 and F_2 respectively. Then the similarity between F_1 and F_2 $d(F_1, F_2) = d(n_1, n_2)$ was defined by the recursive formula below:

$$d(i,j) = \min \begin{cases} d(i-1, j-1) + |\, sf_{1,i} - sf_{2,j}\,| \\ d(i-1, j-2) + \frac{3}{4} \cdot (|\, sf_{1,i} - sf_{2,j-1}\,| + |\, sf_{1,i} - sf_{2,j}\,|) \\ d(i-1, j-3) + \frac{2}{3} \cdot (|\, sf_{1,i} - sf_{2,j-2}\,| + |\, sf_{1,i} - sf_{2,j-1}\,| + |\, sf_{1,i} - sf_{2,j}\,|) \\ d(i-2, j-1) + \frac{3}{4} \cdot (|\, sf_{1,i-1} - sf_{2,j}\,| + |\, sf_{1,i} - sf_{2,j}\,|) \\ d(i-3_i, j-1) + \frac{2}{3} \cdot (|\, sf_{1,i-2} - sf_{2,j}\,| + |\, sf_{1,i-1} - sf_{2,j}\,| + |\, sf_{1,i} - sf_{2,j}\,|) \end{cases}$$

This formula means that the sequences F_1 and F_2 can be matched by "warping" them non-linearly in the time dimension. One syllable from the first sequence can be matched up to three syllables from the second sequence and vice versa.

Distances between all possible pairs of phrases in the corpus need to be calculated using this formula. For this purpose an efficient dynamic time warping procedure (Vintsyuk, 1968) was implemented that estimates phrase-to-phrase distances in polynomial time $O(n_1 * n_2)$

3.2.3 K-means based clustering

Having obtained distances among all possible pairs of phrases, k-means clustering can be applied to them. The clustering consists of iterating through every phrase, removing it from its current cluster and placing it into the nearest cluster (it may also be the same cluster). The distance between the phrase F and the cluster C is estimated by:

$$d(F,C) = \frac{1}{|C|} \sum_{F_i \in C} d(F,F_i)$$

The initial assignment of phrases to clusters is randomized. However the assignment process forbids placing or later moving a phrase into a cluster if there exists at least one other phrase in that cluster that is incompatible with the phrase under scrutiny (the distance between the two might not be calculated due to the required extension/ contraction exceeding the factor of 3). This prevents the formation of mutually incompatible sub-clusters within a single cluster. The k-means algorithm stops when either there's no phrase changing its cluster through the entire iteration or the maximum number of allowed iterations is reached.

4 Experimental evaluation

Human experts have selected 14 male and 9 female speakers (23 hours of speech) out of our 50 hour speech corpus. They selected speakers whose reading intonation was as expressive as possible as well as the accuracy of time aligned transcriptions. There were 41417 spoken phrases in the selected part of the corpus. Spoken phrases had a wide variety of durations.

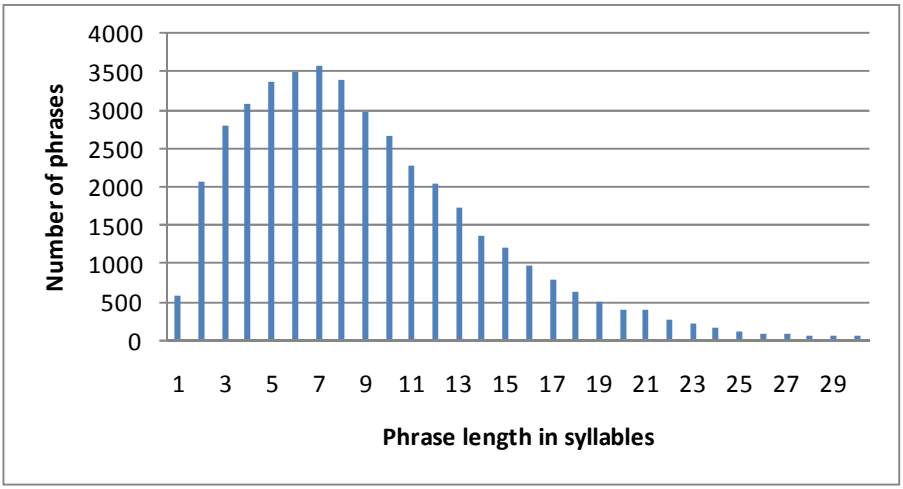

FIGURE 2: The histogram of phrase durations in syllables.

The set of 41417 phrases was divided into two subsets of "short" (15417 items) and "long" (26000 items) phrases[4]. A phrase was considered to be short if it had 6 syllables or less. Otherwise it was considered to be long. The number of clusters *k* was set to 40 and 80 for the sets of short and long phrases respectively. The clustering algorithm was run for a maximum of 100 iterations, though 10 iterations were generally enough for the convergence (see fig. 3).

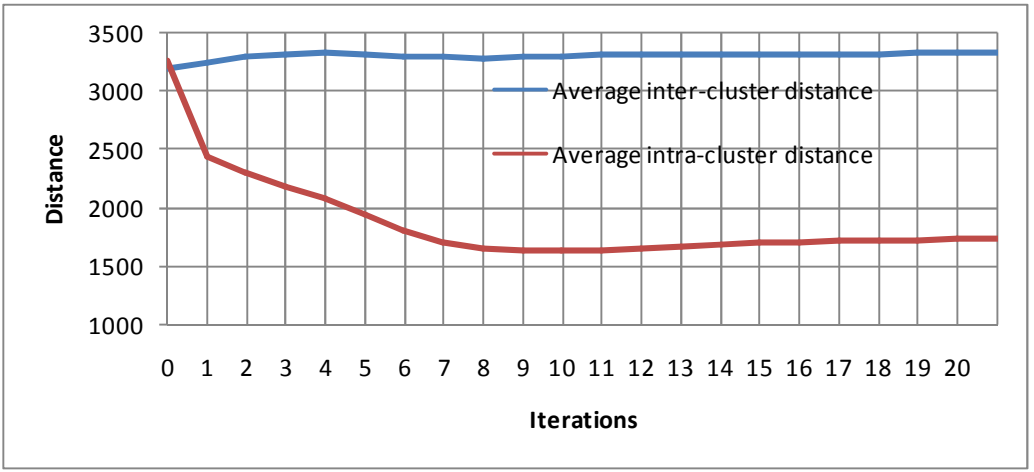

FIGURE 3: The convergence of the k-means clustering algoritm while processing the long phrase set.

Once the clustering process was finished the clustered data was concatenated into one audio and one annotation file per cluster, so that clusters could be visualised (see fig. 4) by Praat software (Boersma and Weenink, 2006).

Human experts complained about clusters being denoted by integer numbers, so we had to look for a more comprehensive naming approach. The decision was taken to find the most "central" phrase of a cluster, i.e. the phrase which has the least average distance to every other phrase in that same cluster, and to use it as the basis for naming the cluster. Every syllable of such "central" phrase was assigned an integer number denoting its pitch in semitones either below (L) or above (H) the normalized pitch average. For instance, the cluster name H5_L5_H1 signifies that the "central" phrase of this cluster is characterized by pitch movements of the high-low-high type that are 5 semitones above, 5 semitones below, and 1 semitone above the normalized pitch average respectively.

[4] This division was due to the limitations in computer memory. It would have been necessary to store 857 mln. distances between all possible pairs of 41417 phrases. We assume that this division is not significant because many phrases from the "short" set and from the "long" set cannot belong to the same cluster simultaneously due to the maximum allowed warping factor of 3.

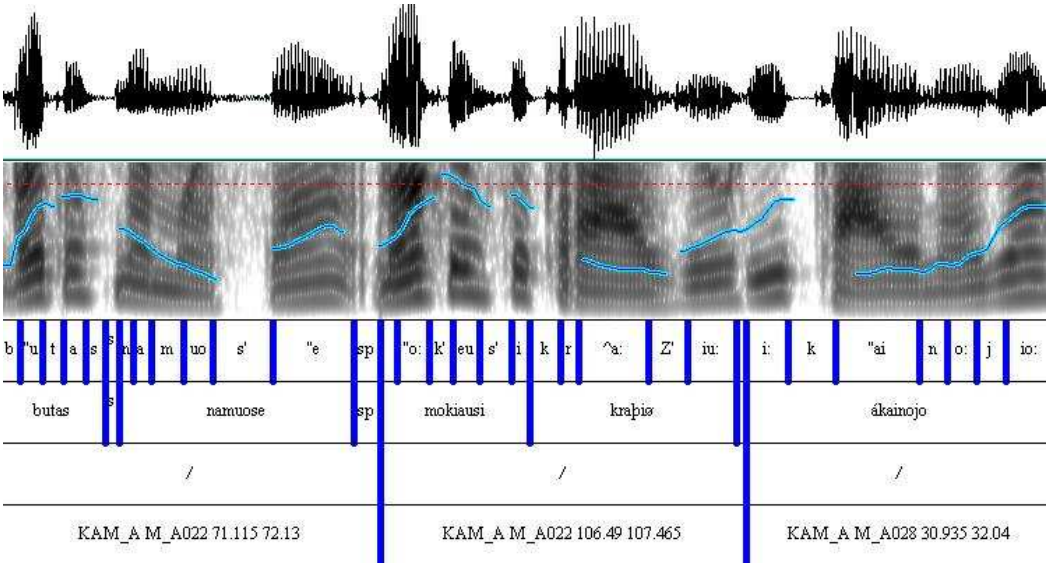

FIGURE 4: An excerpt of the clustering results (cluster H5_L5_H1) visualised by Praat software. Concatenated speech waveform is shown on the top, pitch contours are shown in the middle, and concatenated annotations (phone, word, phrase levels) are at the bottom.

Clustering results are presently being examined by professional phoneticians. Though the evaluation is still in progress, the following remarks can already be made:

- Clustering procedure is efficient in revealing annotation errors (misplaced phrase boundaries). Such phrases tend to form their own clusters.
- Some clusters are dominated by phrases of a few particular speakers. This observation needs further investigation as it may indicate that either pitch normalization is still imperfect or speaker-specific intonation patterns are discovered (e.g. dialect).
- Some clusters contain phrases that should be placed to different clusters from the perceptual point of view. This is mainly due to the emphasis put on different words. This suggests reviewing our present inter-phrase distance metrics and probably including other physical characteristics of speech such as intensity and duration.
- Some clusters appear to be related to certain phenomena of spoken language. There are clusters characterized by the imitation of emotions (not natural for read speech) by the enumerative intonation, by ellipses indicating an unfinished thought at the end of a phrase, etc.

5 Discussion and future work

Language-independence is an important feature of our computational approach. Though computation steps covered by the section 2 are language-specific, all computation steps covered by the section 3 (intonation analysis and clustering) are language-independent.

If speech waveforms and the knowledge of phone and phrase boundaries is available for some language, the set of abovementioned computation steps could be applied.

Computations have been implemented in C++ and optimized for speed. The most costly computational procedure is the step of inter-phrase distance estimation. It has the worst case complexity of $O(N^2*n^2)$, where N denotes the number of phrases in the corpus and n denotes the average phrase length in syllables.

Processing	N	n	Computation time, hours
Short phrases	15417	4.11	1.68
Long phrases	26000	11.87	12.74

TABLE 1: Computation time on a computer with Intel® Core™ 2 Duo T7300 2.00 GHz processor and 3 GB of memory.

The table 1 shows that despite speed optimizations scaling this procedure to larger intonation corpora may be problematic. However, we believe that the discovery of intonation patterns proper to a given language could be made within smaller but more carefully selected intonation corpora.

Our future research will be mostly driven by the feedback received from human experts. In addition to those few research issues that have been mentioned in the section 4.2, our future research may focus on the optimization of the number of clusters, on different methods of estimating the representative frequency of a syllable, on the normalization of this frequency with respect to the intrinsic vowel pitch.

Acknowledgements

This research has been supported by a grant from the Research Council of Lithuania under the National Lithuanian studies development programme for 2009-2015. through the project A unified approach to Lithuanian prosody: the intonation, rhythm, and stress (reg. no. LIT-12020).

References

Beckman, M. E., Hirschberg, J., and Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (ed.) *Prosodic Typology -- The Phonology of Intonation and Phrasing.*

Boersma, P., and Weenink, D. (2006). Praat: doing phonetics by computer [Computer program]. Version 4.4.26, retrieved 24 July 2006 from `http://www.praat.org/`

Hartigan, J. A.; Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society*, Series C (Applied Statistics) 28(1):100–108.

Kundrotas, G. (2008). Lietuvių kalbos intonacinių kontūrų fonetiniai požymiai (Phonetic features of Lithuanian intonation contours). *Žmogus ir žodis. Didaktinė lingvistika*. Mokslo darbai, Vilnius, 10(1): 43-55.

Kundrotas, G. (2009). *Lyginamoji lietuvių ir rusų kalbų intonacinių sistemų analizė* (A comparative analysis of intonation systems of Lithuanian and Russian), Vilnius Pedagogical University.

Levow, G.-A. (2008). Automatic Prosodic Labeling with Conditional Random Fields and Rich Acoustic Features. In *Proceedings of International Joint Conference on Natural Language Processing (IJCNLP)*, pages. 217-224.

Heggtveit, P. O., Natvig, J. E. (2004). Automatic prosody labelling of read Norwegian. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, vol. 4, pages 2741-2744.

Escudero-Mancebo, D., Vizcaíno-Ortega, F., González-Ferreras, C., Vivaracho-Pascual, C., Cabrera-Abreu, M., Estebas-Vilaplana, E., and Valenítn Cardeñoso-Payo, V. (2012). Multiclass Pitch Accent Classification for Assisting Manual Prosodic Labeling. In *Proceedings of IberSPEECH 2012*, pages 73-82.

Norkevičius, G., Raškinis, G., and Kazlauskienė, A. (2005). Knowledge-based grapheme-to-phoneme conversion of Lithuanian words. In *Proceedings of the 10th International Conference on Speech and Computer – Specom*, Patras, Greece. pages 235–238.

Raškinis, G. (2000). Lietuvių liaudies dainų užrašymas muzikos simbolių sekomis (Automatic transcription of Lithuanian folk songs), Phd Thesis, Vytautas Magnus University.

Vaičiūnas, A. (2006). Lietuvių kalbos statistinių modelių ir jų taikymo šnekos atpažinimui tyrimas, kai naudojami labai dideli žodynai (Investigation of Lithuanian statistic language models and of their application to speech recognition in case of very large vocabularies), Phd Thesis, Vytautas Magnus University.

Vereecken, H., Martens, J.-P., Grover, C., Fackrell, J., and Van Coile, B. (1998). Automatic prosodic labeling of 6 languages. In *Proceedings of the 5th International Conference on Spoken Language Processing*, Sidney Australia.

Vintsyuk, T. K. (1968). Speech discrimination by dynamic programming. *Kibernetika*, 4:81-88.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2002). *The HTK Book* (for HTK Version 3.2), Microsoft Corporation, Cambridge University Engineering Department, pages 22-44.

Xu, Y. (2004). Transmitting tone and intonation simultaneously - the parallel encoding and target approximation (PENTA) model. In *Proceedings of the International Symposium on Tonal Aspects of Languages (TAL) - 2004*, pages 215–220.

Wightman, C., and Ostendorf, M. (1994) Automatic labeling of prosodic patterns, *IEEE Transactions on Speech and Audio Processing*, 2(4):469–481.