A Speech to Text Transcription Approach based on Romanian Corpus

Andrei Scutelnicu
Faculty of Computer Science, Alexandru Ioan Cuza University of Iasi, Romania
16 General Henri Mathias Berthelot, Iași 700259, Tel.: +40 232 201 090
Institute of Computer Science, Romanian Academy, Iași, Romania
Romanian Academy, Codrescu, Iași 700481, Tel. +40 332 106 505
andreiscutelnicu@gmail.com

Mihaela Onofrei
Faculty of Computer Science, Alexandru Ioan Cuza University of Iasi, Romania
16 General Henri Mathias Berthelot, Iași 700259, Tel.: +40 232 201 090
Institute of Computer Science, Romanian Academy, Iași, Romania
Romanian Academy, 2 Codrescu, Iași 700481, Tel. +40 332 106 505
mihaela.plamada.onofrei@gmail.com

Anca Diana Bibiri
Alexandru Ioan Cuza University of Iasi, Romania
Bulevardul Carol I 11, Iași 700506, Tel. +40 232 201 000
anca.bibiri@gmail.com

Mircea Hulea
Faculty of Automatic Control and Computer Engineering, Gheorghe Asachi University of Iași,
Romania, 67 Bulevardul Profesor Dimitrie Mangeron, Iași 700050, Tel. +40 232 278 683
mhulea@tuiasi.ro

Abstract

Automatic speech segmentation has many applications in speech processing and phonetics, e.g., in automatic speech recognition and automatic annotation of speech corpora. In both processes of training and evaluation of speech recognition systems large aligned speech-to-text corpora are needed. Once aligned, identification of phonemes could be based on samples that are picked-up in-between phonemes' boundaries. Because manual segmentation is costly and extremely time consuming, automatic methods of alignment are searched for. In this paper, we propose a simple, yet efficient, method for speech to text recognition based on a machine learning approach, using a Romanian speech corpus.

Keywords: speech recognition, formant energy, algorithmic method, neural network;

1. Introduction

Automatic speech segmentation is suitable for speech to text conversion and might improve the speech recognition results. In this paper, we propose a method of recognition of the speech signal based on analyses of variations in the energy of formants. The method that we have used in this paper for speech to text recognition is based on neural networks algorithm.

2. Work In Speech Recognition

The goal of an automatic speech recognition (ASR) system is to find, given an acoustic signal (an utterance), its most probable corresponding words sequence. The ASR term is also known as speech-to-text transcription, having application in human-computer interaction such as:

- hand-busy or eye-busy applications;
- telephony;
- transcription of courses, monologue of a speaker, courthouse;
- augmentative communication for human with inability to type.

The problem of Large-Continuous Speech Recognition (LVCSR) is addressed in transcription systems with vocabulary of 20.000 to 64.000 words (Jurafski et al. 2006, ch. 9; Hunag et al. 2007).

Most systems make use of statistical models (Huang 2001, Jurafsky 2006, Jelinek 1976). From this point of view, the speech-to-text task can be formulated mathematically as follows:

$$W^* = \arg\max_{W \in \mathrm{W}} P(W \mid A) \quad (1)$$

where W* ∈ W is the word string and A is the observed acoustics.

Using Bayes decomposition rule, and knowing that the input acoustics are fixed and given to us, the above function can be rewritten as:

$$W^* = \underbrace{\underset{W \in \mathbf{W}}{\mathbf{argmax}}}_{\text{search}} \underbrace{P(A|W)}_{\text{AM}} \underbrace{P(W)}_{\text{LM}} \quad (2)$$

where the first probability distribution, P(A|W), is the observation likelihood and becomes the characterization of an acoustic model (AM), while the second probability distribution, P(W), the a priori probability, becomes the characterization of a language model (LM). The optimal hypothesis is found out by searching (argmax) through all word sequences possible in the language.

Figure 1 shows the main components of a speech recognition system:

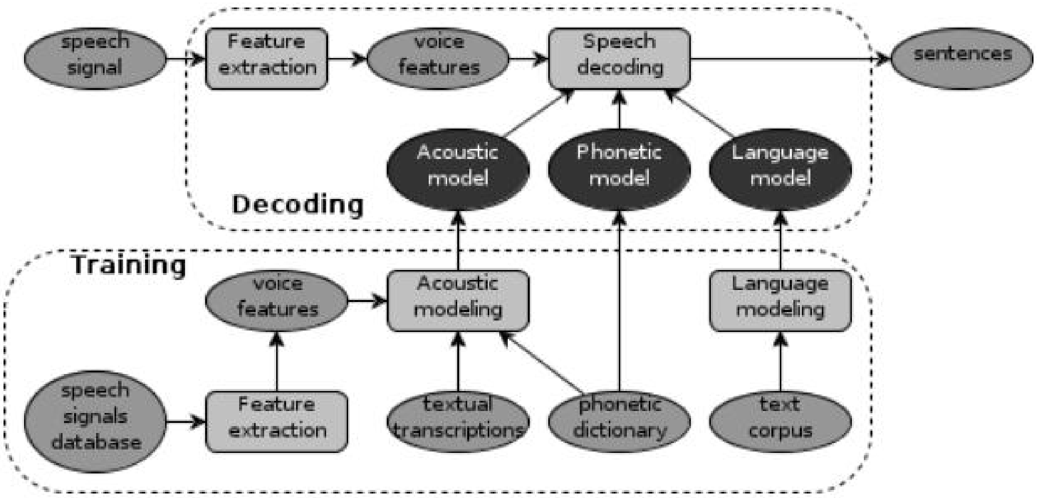

Figure 1. Main components of an ASR system (Cucu, PhD. Thesis, 2011)

A speech recognition system has three main sources of information: a speech signals corpus, a text corpus and a phonetic dictionary.

The recognition has the next main phases:

- signal processing phase, in which a vector of around 42 features (39 of spectral information, energy, spectral change) for each time window (Jurafski et al., 2006; Hunag et al., 2007);
- acoustic modeling or phone recognition phase, the estimation of P(W|A), for each linguistic unit is computed a likelihood of the observed feature vectors; include the representation of knowledge about acoustics, phonetics, microphone and environment variability, gender and dialect differences among speakers, etc.

- language modeling phase, the estimation of P(W), refer to a system's knowledge of what constitutes a possible word and in what sequence,
- decoding/searching phase of a speech recognizer, where the most likely word is given based on the sequence of acoustic likelihoods, a phonetic dictionary of word pronunciations and a language model.

An acoustic model contains statistical representations of each of the distinct sounds (called phonemes), that makes up a word in a language. The Romanian language has about 38 distinct sounds (Chiţoran, 2001; AR, 2005) that are useful for speech recognition, and thus we have 38 different phonemes. An acoustic model is created by taking a large database of speech and using special training algorithms to create statistical representations for each phoneme in a language. These statistical representations are called Hidden Markov Models (HMMs) (Baker, 1975; Poritz, 1988; Rabiner, 1989; Jelinek, 1998). Each phoneme has its own HMM.

HMMs are probabilistic finite state machines, which may be combined hierarchically to construct word sequence models out of smaller units. In large-vocabulary speech recognition systems, word sequence models are constructed from word models, which in turn are constructed from sub-word models (typically context-dependent phone models) using a pronunciation dictionary.

A phonetic model is usually a pronunciation dictionary that maps all the words in the vocabulary to a sequence of phones. The phonetic dictionary can be regarded as an interface between the acoustic model, which works with phones and the language model which works with words.

The development of a phonetic dictionary is an important, but difficult task. The task is time-consuming and tedious (for a manual development) and also requires a very good knowledge of the language.

The problem of Large-Continuous Speech Recognition (LVCSR) is addressed in transcription systems with vocabulary of 20.000 to 64.000 words [Jurafski et al., 2006, ch. 9; Hunag et al., 2007].

3. The Method

The activity of collecting of a medium-size speech-to-text aligned corpus (Bullinaria, 2011), allowed a visual inspection of the energy of the formants at the segments' boundaries. Figure 2 shows the energy of the formants F1, F2, and F3, for the word *pasărea* (the bird). Abrupt changes can be noticed in the energies of formants at phone boundaries and relatively stable shapes in-between boundaries. The algorithm that we describe below is inspired by these findings.

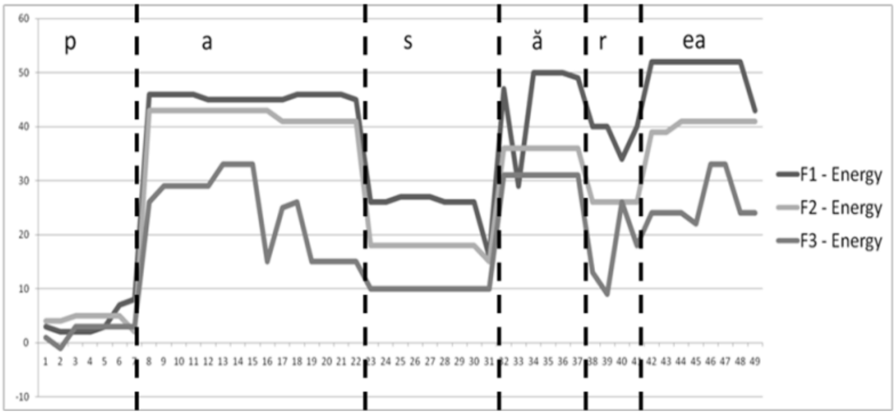

Figure 2. Formants' energy for the word pasărea (bird)

After examining many such diagrams, it became evident that the energy of the F2 formant seems to have the most stable behavior with respect to the abrupt changes at the phones boundaries and to calmness in the interior of phones. Ulterior, quantitative evaluations confirmed the claim that the energy of F2 can be reliably used as supporting a rule-based speech to text recognition algorithm.

Based on these observations, an automatic segmentation algorithm was carried out, based on the energy values of F2 as follows:

- groups of 3 values were averaged;
- values were extracted by means of mediations result;

We imposed the following rules:

- Test 1:
    - if the value v [i + 1] is greater than the value v [i] then: we calculated the average of two consecutive values v [i] and v [i + 1], which must be greater than v [i] and the difference between v [i + 1] v [i] be greater than a threshold imposed as a parameter (chosen by empirical observations to THRESHOLD1 = 8;

```
if (v[i+1] - v[i] > THRESHOLD1)     (1)
```

    - if the value v [i + 1] is lower than the v [i] then: we calculated the average of two consecutive values v [i] and v [i + 1], which must be greater than v [i] and the difference between v [i] and v [i + 1] is greater than a threshold imposed as a parameter (THRESHOLD2 = 5):

```
if (v [i] - v [i + 1]> THRESHOLD2) (2)
```

- Test 2:
    - if the value v [i + 1] is greater than the value v [i] then: we calculated the average of two consecutive values v [i] and v [i + 1], which must be greater than v [i] and the difference between v [i + 1] v [i] must be greater than a threshold imposed as a parameter (chosen by empirical observations to THRESHOLD1 = average of all inputs / 4)

```
if (v [i + 1] - v [i]> = THRESHOLD1) (1)
```

    - if the value v [i + 1] is lower than the v [i] then: we calculated the average of two consecutive values v [i] and v [i + 1], which must be greater than v [i] and the difference between v [i] and v [i + 1] is greater than a threshold imposed as a parameter (THRESHOLD2 = average of all entries / 5):

```
if (v [i] - v [i + 1]> THRESHOLD2) (2)
```

For automatic recognition of phonemes, we used a neural network algorithm - *Feed Forward - Back Propagation* (Figure 3).

The neural network comprises an input layer - X (Figure 3 to 9 neurons), a hidden layer of neurons and an output layer - Y (in Figure 1 to 6 neurons). Network weights are chosen randomly and are stored in a three-dimensional matrix, where the transmitting neuron *i* represents index and *j* is the index of the receiving neuron; It is the index of the layer. Choose a set of input data X (input vector values) and a data output Y (output vector values) representing the desired outputs. Each neuron in the hidden layers, respectively, of the output layer performs a weighted summation of the inputs and applying activation functions of these amounts. If the perceptron shown in Figure 3, the neurons in the output layer performs a weighted sum of:

$$s_j = \sum_i x_i w_{ij} \text{ , for } i=\overline{1,n} \text{ ; } j=\overline{1,m} \text{ ,}$$

n - input neurons,
m - output neurons;

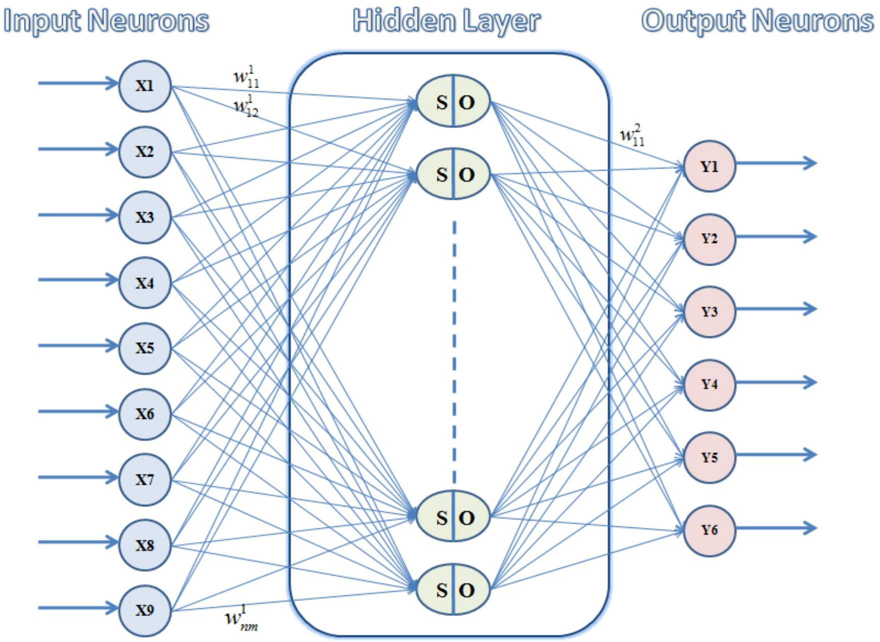

Figure 3. Feedforward Back Propagation - neural network

After calculating the amounts s_j were calculated neural networks outputs $y_j = f(s_j)$, where $f(s_j)$ is the activation function of neurons. Activation function used in the implementation of the algorithm presented in this paper is sigmoidal function:

$$y_j = \frac{1}{1+e^{-s_j}} \text{ , } j=\overline{1,m} \text{ .}$$

For each pair of vectors (X_p, Y_p) are executed the following steps:

- Calculate O_i generated by the network output, which is obtained by forward propagation of the input vector X_p, $p=\overline{1,N}$;
- Calculate the error output generated compared to desired output $E_p = \frac{1}{2} \cdot \sum_{j=1}^{n} (y_j - o_j)^2$, where $y_j \in Y_p$, $o_j \in O_p$, $p=\overline{1,N}$, and *n* represents the output number of the neural network;
- Calculate the error gradient: $\nabla E_p = \left\{ \frac{dE_p}{dw_{1j}}, \frac{dE_p}{dw_{2j}}, \frac{dE_p}{dw_{3j}}, ..., \frac{dE_p}{dw_{mj}} \right\}$, where w_{ij} = $i=\overline{1,m}$ represents weights to neuron input x_i to neuron *j*, belonging to the output layer.
- Adjust output layer weights purpose of minimizing error, as follows:

  $$w_{ij} = w_{ij} + \Delta w_{ij} \text{ , where } j=\overline{1,n} \text{ , } i=\overline{1,m} \text{ ; (1)}$$
- Variation weights will be calculated based on the output neurons of the previous layer and the gradient of the error:

$$\Delta w_{ij} = -\alpha \cdot x_i \frac{dE_p}{dw_{ij}} \quad (2)$$

- The neuron activation function is $g(s_j) = \frac{1}{1+e^{-s_j}}$, and integration function is $s_j = \sum_{i=1}^{m} w_i x_i$; derived error is calculated as follows:

$$\frac{dE_p}{dw_j} = \frac{d\left(\frac{1}{2}\cdot\sum_{j=1}^{n}\left(y_j - g(s_j)\right)^2\right)}{dw_j} \quad (3)$$

- After replacing the equation (3) in (2) we are calculating new weights: $w_{ij} = w_{ij} + \Delta w_{ij}$;
- After performing a calculation according to (2) we have:

$$\frac{dE_p}{dw_{ij}} = o_j(1-o_j)(y_j - o_j), \ j = \overline{1,n}, \ i = \overline{1,m}$$

$$w_{ij} = w_{ij} + \alpha \cdot x_i \frac{dE_p}{dw_{ij}} \quad (4)$$

  - As a vector output for hidden layers which aims to be achieved is not known, the gradient output is calculated as:

$$\frac{dE_p}{dw_{(i-1)i}} = o_i(1-o_i)(\sum_{r=1}^{m}\frac{dE_p}{dw_{rj}} \cdot w_{rj}) \quad (5),$$

    where *j* is the output layer, and *i* and *(i - 1)* the hidden layer.
  - New weightings for the hidden layer (i - 1) adjusted using equality (4), followed by calculation error gradient layer (i - 1) using equation (5). Similarly, using expressions (4) and (5) adjust the weights for all hidden layers of the network.

4. Analyzed data

The data used for the segmentation is based on the two questionnaires (AMPER-ROM[ANIA] and AMPRom) used in the project AMPRom (Romanian multimedia prosodic atlas). This is the first prosodic atlas which aims to present the main prosodic patterns (intonation patterns) of the Romanian language varieties identified both at the level of the diatopic variants of the standard language and at the level of the dialect variants. The data was collected during the dialectal prosodic surveys in the network of points that covers the whole territory of the Dacoromanian dialect.

The sets of statements, established by morpho-syntactic and phonetic criteria, are formed by: declarative sentences (affirmative and negative) and total interrogative sentences (affirmative and negative), having the syntactic structure of SVO (subject - verb - object) where S and O receive, in turns, adjective and / or prepositional determinates; the words of the sentences follow the type CV (consonant + vowel) formed of voiceless consonant + (semi)open vowel, which facilitates the segmentation of the phonetic elements and the 45 sentences of AMPER-ROM[ANIA] questionnaire are made up of trisyllabic nouns: căpitan -captain- (oxiton), nevasta -wife- (paroxiton) and pasărea-bird-(proparoxiton), seven trisyllabic adjectival determinants: frumoasă -beautiful-, harnică -diligent-, tinerea -young-, galbenă -yellow-; elegant -elegant-, amabil -kind-, repede -fast- and one noun determinant: (pasărea -bird-) papagal -parrot- and a verbal form vede -see- (when the noun captain is a direct object of the verb, usually, syneresis – contraction of two vowels into a diphthong – is produced as a feature of spontaneous speech: vede-un căpitan -sees a captain-). This date is used for testing, while for training the data is slightly different and consist of a large number of other types of utterances (those from AMPRom corpus).

Although the manual segmentation and labelling of the speech corpus constitutes a more accurate method (the segmentation accuracy may vary according to the different human phonetic expert criteria), this task requires a large amount of time and concentration for the human labeller.

An accurate segmentation that would be done fully manually would require as many as 800 times real-time, i.e. 13 hours for a one-minute recording (Schiel 2003). The processing time is a major drawback for manual labelling, especially when faced with very large spontaneous speech corpora. Thus, an automatic phonetic alignment tool is highly desirable.

Acoustic analysis tools that are used in processing the prosodic dialectal data that was recording during the investigations are PRAAT/SCRIPT PRAAT for AMPER (Antonio Romano, Albert Rilliard), Matlab, AMPER 2006, Computer interface of prosodic curve. Statements are recorded in digital format (files with .wav extension - Waveform Audio File Format) and acoustic analysis using software tools. The sequence analysis goes through several stages: changing the sampling frequency sound wave of 48 kHz to 16 kHz (Gold Wave) delineation and labelling according to the statements used in the questionnaire.

The PRAAT software was used to display oscillograms and spectrograms of the speech signal. By visualizing these channels and simultaneously hearing the sound, the human expert placed segmentation boundaries and labels each segment with the corresponding phonetic element (in the case of diphthongs the two vowels go together). Thus, for each analyzed utterance, physical correlates of vowels (duration, intensity and fundamental frequency – F0, for the three points of the vowel) are recorded. The annotation levels in the corpus are: utterance, word, syllable, phoneme and grapheme. Matlab routines have been used then to compute average values, duration graphs, intensity and individual melodic lines.

In order to extract as many features needed for automatic alignment several tiers were generated: one tier with phones and another with words. Additionally, a syllable tier is generated on the basis of sonority-based rules for syllable segmentation. In summary, the whole manual process results in a multi-level alignment annotation between speech and text within a TextGrid composed of phonetic, syllabic, lexical and, optionally, utterance tiers.

5. Results and future work

Following the implementation of the algorithm, we conducted the following tests:

- Firstly, we considered the distinction vocal - consonant we have made some choices: for the input neurons, we have selected the values for formant F0 and amplitude A0; for the output the signal was binary encoded as 1- for vocal, 0 for consonant; the number of iteration for training the neural network were as follow 100, 1000 and 3000 iterations. For this test the recognition algorithm recognition rate was between 54% and 56%;
- In the second test, we have taken values for the input neurons for formants F0, F1, F2, F3 and amplitude A0; for training the neural network we have taken 100, 2000, 5000 iterations and the recognition result was between 62% and 68%.

There is an improvement in outcome with an 8% -> 12% after using the 3 values and higher forms.

The results presented in this paper are at first attempt. The next step is to add more features and to test the method on a greater variety of data such as increasing the training corpus, test it on more input values for neural network, improving the results with a statistical model and a language model.

Acknowledgement

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS-UEFISCDI, project number PN-II-RU-TE-2014-4-2298.

References

Andre-Obrecht, R. (1986). Automatic Segmentation of Continuous Speech Signals, Proc. of ICASSP-Tokyo, 2275-2278

Bibiri, A., Cristea, D., Pistol, L., Scutelnicu, L., & Turculeţ, A. (2013) Romanian Corpus For Speech-To-Text Alignment, Proceedings of the Workshop on Linguistic Resources and Instruments for Romanian Language Processing, ConsILR-2013, Miclăuşeni - Iaşi, The Editorial House of ”Al. I. Cuza” University, Iaşi, ISSN 1843-911X, 151 -162

Boersma, P., & Weenink, D. (2010). Praat: doing phonetics by computer, http://www.praat.org

Bullinaria, J. A. (2011). Text to Phoneme Alignment and Mapping for Speech Technology: A Neural Networks Approach, IJCNN, IEEE, 625-632

Cole, R., & Hou, L. (1988). Segmentation and broad classification of continuous speech. Proceedings of ICASSP, 453-456

Gómez, J. A., Sanchis, E., & Castro-Bleda, M. J. (2010). Automatic Speech Segmentation Based on Acoustical Clustering, Proceedings of Structural, Syntactic, and Statistical Pattern Recognition, Joint IAPR International Workshop, SSPR & SPR 2010, Cesme, Izmir, Turkey, Vol. 6218, 540-548

Glass, J.R., & Zue, V.W. (1988). Multi-level acoustic segmentation of continuous speech, Proceedings of ICASSP, 429-432

Hatazaki, K., Komori, Y., Kawabata, T., & Shikano, K. (1989). Phoneme segmentation using spectogram reading knowledge. In Proceedings of ICASSP, pages 393-396, 1989.

Juang, B. H., & Rabiner, L. R. (1990). The Segmental K-means algorithm for estimating parameters of Hidden Markov models, IEEE Trans. on Acoustics, Speech and Signal proc., 38(9):1639-1641

Ljolje, A., & Riley, M. D. (1991). Automatic segmentation and labeling of speech, Proceedings of ICASSP, 473-476

Reddy, D. R. (1966). Segmentation of Speech Sounds, J.Acoust.Soc.Am, Vol. 40, Issue 2, 307-312

Schiel, F., & Draxler, C. (2003). The Production of Speech Corpora. Munich: Bavarian Archive for Speech Signal

Svendsen, T., & Soong, F. K. (1987). On the Automatic Segmentation of Speech Signals, Proceedings of ICASSP - Dallas, 77-80 & 341-344

Toledano, D. T., Hernandez Gomez, L. A., & Grande, L. V. (2003). Automatic Phonetic Segmentation, IEEE Trans. Speech and Audio Proc., Vol 11, Issue 6, 617-625

Van Hemert, J. P. (1991). Automatic Segmentation of Speech, IEEE Trans. on Signal Proc., Vol. 39, Issue 4, 1008-1012