Segmentation of Conversational Speech Using Probabilistic Neural Network

Dr. Ahmed Maamoon Alkababji
Lecturer

Computer Engineering Department, Collage of Engineering,
University of Mosul, Mosul, Iraq.

Abstract

Automatic segmentation of audio streams according to speaker identities, environmental and channel conditions has become an important preprocessing step for speech processing, speaker recognition and audio mining. This paper presents an automatic speech segmentation system where the performance of the probabilistic neural network (PNN)(which is the main part of the system) is examined and then enhanced in the area of segmentation of conversational speech. The results show that a percentage false segmentation (PFS) of 18% can be achieved. PFS is dropped to 6.1% enhancing the system. The experiments were carried out on a dataset created by concatenating speakers from the TIMIT database.

Keywords: Speech segmentation, PNN, Probabilistic neural network.

تقطيع المحاورات الكلامية باستخدام الشبكة العصبية الاحتمالية

د.احمد مأمون فاضل
مدرس
قسم هندسة الحاسبات, كلية الهندسة, جامعة الموصل, الموصل, العراق

الملخص

لقد أصبح التقطيع الذاتي للاسترسال الكلامي اعتماداً على هوية المتكلم أوالبيئة التي تم الكلام فيها أو القناة التي نقلت الكلام من المراحل المهمة في المعالجة المسبقة التي تتم على الكلام لعمليات مثل تمييز المتكلم أو تمييز الكلام. في هذا البحث تم تقديم نظام تقطيع ذاتي تم من خلاله استكشاف إمكانيات الشبكة العصبية الاحتمالية على تقطيع المحاورات الكلامية كما تم تحسين أداء النظام للحصول على نتائج أفضل حيث أظهرت النتائج إمكانية الحصول على نسبة خطأ في التقطيع مقدارها 18%, في حين انخفضت هذه النسبة إلى 6,1% بعد إجراء عملية التحسين على النظام. تم استخدام مجموعة TIMIT كقاعدة أصوات لتقييم أداء النظام.

Received 15/12/2008 **Accepted 18/8/2009**

1.Introduction

Automatic speech segmentation aims to find the speaker change points in an audio stream. It is a preprocessing task for audio indexing, speaker identification - verification - tracking, automatic transcription, information extraction, topic detection, speech summarization and retrieval[1].

Several techniques have been used for speech segmentation. Among them are those based on Bayesian Information Criterion(BIC) [1]. In [2] the (BIC) is compared with the Cumulative Sum (CuSum) algorithm for automatic segmentation. The Use of an adaptive Vowel/ Constant/ Pause (V/C/P) classification method [3] attempt to segment speech without speech recognition. The use of Artificial Neural Network (ANN) was for more than a decade as in [4] for automatic speech segmentation and its performance was compared to that of Hidden Markov Models (HMMs).

In the present work a special type of neural network is used. This neural network uses a kernel-based approximation to form an estimation of the Probability Density Functions (PDFs) of categories in a classification problem; this is the Probabilistic Neural Network (PNN). This particular type of ANN provides a general solution to pattern classification problem by following the probabilistic approach based on the Bayes decision theory [5].

The rest of the paper is organized as follows. The overview of our system including feature extraction is described in Section 2. Section 3 presents a brief description of the data base used for the evaluation of the proposed segmentation system. Section 4. demonstrates the heart of the proposed system (PNN) and its training as well as the results of testing the system by the selected data base. In addition to that a new technique is described and tested on the same data base showing an advancement in the segmentation results of the proposed system. Finally, in section 5, some concluding remarks of this work is given.

2.System Overview

The proposed segmentation system relays on the neural network to make the decision to which speaker, a currently examined segment belong. Thus the system has two phases; the training phase and the testing or segmentation phase. In both phases the speech sample must be preprocessed. The preprocessing includes framing, windowing, feature extraction. This is achieved by the stages described in the following subsection. Figure 1 shows the phases of the system and the stages of each phase.

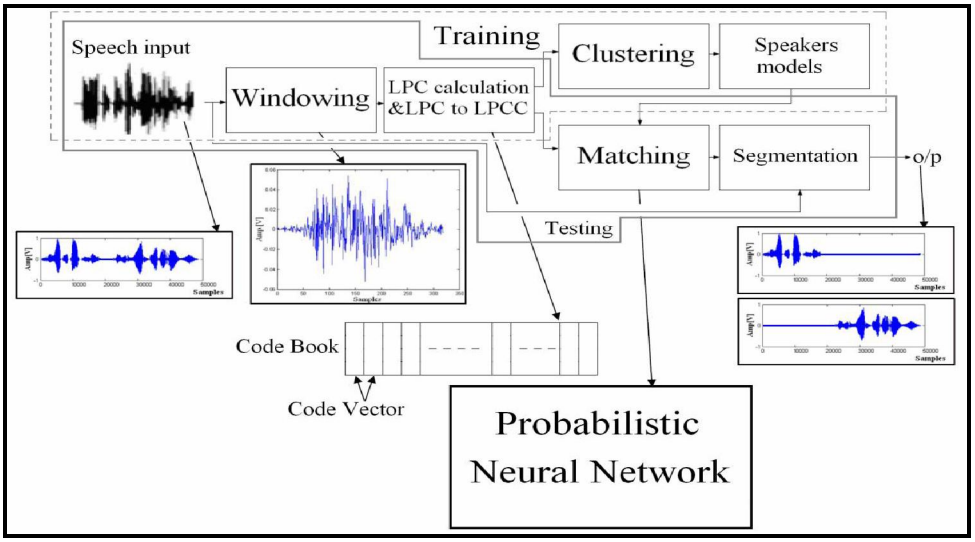

Figure 1: The phases and stages of the proposed segmentation system.

2.1 Framing and Windowing

This is the front end of the system and the first stage of the preprocessing that is to be carried out to prepare the input speech signal to the next stage (the feature extraction stage). In this stage the speech signal is segmented into 20ms long frames with an overlap of 10ms. Then each frame is multiplied by a window function to reduce the effect of the spectral artifacts that result from the framing process. The Hamming window is used in this system for such aim.

The selection of the frame length is a crucial parameter for successful spectral analysis, due to the trade-off between the time and frequency resolutions. The window should be long enough to adequate frequency resolution, but on the other hand, it should be short enough so that it would capture the local spectral properties. Typically, a frame length of 10-30 milliseconds is used [6]. Usually adjacent frames are overlapping by some amount. A typical frame overlap is around 30 to 50 % of the frame size. The purpose of the overlapping analysis is that each speech sound of the input sequence would be approximately centered at some frame [7, 8].

2.2 Feature extraction stage

In the feature extraction stage Linear Predictive Coding (LPC) is used. The model based representation of speech gives rise to Linear Prediction Coefficients model (LPC). LPC is a very important spectral estimation technique because it provides an estimate of the pole of the vocal tract transfer function. The LPC algorithm is an nth order predictor which attempts to predict the value of any point in a time varying linear system based on the values of the previous n samples. The rationale in linear prediction (LP) analysis is that adjacent samples of the speech waveform are highly correlated and thus, the signal behavior can be predicted to certain extent based on the past samples. According to [9] an LPC predictor larger than 15 is sufficient to represent the features of a speech segment. Therefore a 16 LPC predictor is used. This 16 LP coefficients c[n]are then converted to its corresponding LPC Cepstrum coefficients (LPCC) using [10]

$$c[n] = \begin{cases} a[n] + \sum_{k=1}^{n-1} \frac{k}{n} c[k]\, a[n-k], & 1 \le n \le p \\ \sum_{k=n-p}^{n-1} \frac{k}{n} c[k]\, a[n-k], & n > p \end{cases} \quad \ldots.(1)$$

A noticeable thing is that although there are finite number (p) of LP coefficients, the LPC cepstrum sequence c[n] is infinite. However, the magnitudes of $|c[n]| \rightarrow 0$ fast with n, and thus relatively small number of coefficients is needed to model the spectrum[10].

At this point each 20ms frame of the input signal is represented by a vector of 16 LPCC coefficients this vector is called *code vector*. Due to the 50% (10ms) overlap the number of code vectors (N) for a speech signal of duration D can be calculated as follows:

Number of code vectors (N)= [(D-mod(D,10ms))/10ms]-1 ….(2)

For an example, for a 1 second input speech signal there is (1000/20=50) 20ms frames. Due to the overlap the number of frames is multiplied by 2. Therefore, to find the number of frames which equals the number of code vectors, the duration of the speech signal is divided by 10ms. Before division the duration is made a precise multiples of 10ms. It must be noticed that the last frame can not be overlapped. Therefore, the total number of frames (code vectors) must be decreased by 1.

Therefore, the input speech signal is represented by an 16*N matrix (16 LPCC coefficients by N code vectors), this matrix is called *codebook*. Refer to Figure 1 which abstract the derivations of the codebook from the speech signal.

3. Speech Corpus

The speech corpus is used to examine the performance of this proposed speaker segmentation system. It is the standard American English TIMIT provided by Linguistic Data Consortium [11]. TIMIT is an acoustic-phonetic database including 6300 sentences and 630 speakers who speak English. The audio format is PCM, the audio samples are quantized in 16 bit, the recordings are single-channel, the mean duration is 3.28 sec and the standard deviation (st. dev.) is 1.52sec. From all the available data in the TIMIT corpus two arbitrary subsets of speakers are used in this work. The male speaker's subset contained 70 speakers and the female speaker's subset contained 70 speakers too. There are 10 speech files for each speaker; two of the files have the same linguistic content for all speakers, whereas the remaining eight files are phonetically diverse.

For each speaker, a codebook is built using the following process: Three of the ten files available for each speaker are used including the two of the phonetically identical file. As in section 2.1 and 2.2 for each file, a codebook is created. The three codebooks are pooled resulting one large codebook, then the k-mean clustering algorithm was used to cluster this large codebook to obtain an (16*128) codebook (an overall sum of 140 codebooks, 70 for the male speakers and 70 for the female speakers). These codebooks are then used to train the conversation segmentation system.

In a conversation there are three probabilities for the speakers participating in it: a male-male conversation, female-female conversation and male-female conversation. For the testing phase of the conversation segmentation system 70 conversations are created for each of the three probabilities mentioned above by concatenating the remaining seven speech files of the speakers in the selected subsets (a total of 210 conversations). For an example a conversation between speaker 1 and speaker 2 is created as follows:((1st file of speaker 1,1st file of speaker 2,2nd file of speaker 1, 2nd file of speaker 2…7th file of speaker 2).

4. The Probabilistic Neural Network

The main stage of this conversation segmentation system is the probabilistic neural network. This neural network has been used in many speaker recognition systems as in [12-14]. In this work the performance of the probabilistic neural network as a conversation segmentation tool is to be investigated.

A useful interpretation of the network outputs under certain circumstances is to estimate the probability of class membership, in which case the network is actually learning to estimate a probability density function (PDF). This is the case of the probabilistic neural network (PNN).

The network paradigm basically uses the Parzen-Cacoulos estimator to obtain the corresponding PDF of the classification categories. PNN uses a supervised training set to develop probability density functions within a pattern layer [5]. The PNN implements the Parzen window estimator by using a mixture of Gaussian basis functions. If a PNN for classification in K classes is considered, the probability density function $f_i(x_p)$ of each class K_i is defined by:

$$f_i(\boldsymbol{x}_p) = \frac{1}{(2\pi)^{d/2}\sigma^d}\frac{1}{MM_i}\sum_{j=1}^{MM_i}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{x}_p - \boldsymbol{x}_{ij})^T(\boldsymbol{x}_p - \boldsymbol{x}_{ij})\right), i = 1,2,...,K \quad \ldots.(3)$$

where x_{ij} is the j-th training vector from class K_i , x_p is the p-th input vector, d is the dimension of the speech feature vectors, and MM_i is the number of training patterns in class

K_i. Each training vector x_{ij} is assumed to be a centre of a kernel function, and consequently the number of pattern units in the first hidden layer of the neural network is given as a sum of the pattern units for all the classes. The variance σ acts as a smoothing factor, which softens the surface defined by the multiple Gaussian functions.

For each conversation, which in this work contains the speech of two speaker only, a PNN is designed to decide if the input segment (10 msec. long) belongs to the first or to the second speaker. Both speakers are represented by codebooks which are used in training the network. Therefore the problem is reduced to classifying the input test vector to one of two classes (K=2).Figure 2 shows the architecture of the probabilistic neural network used in this segmentation system, the two hidden-layers of the PNN used in this system are shown [15]. The Radial Basis layer is defined as

$$a_i1 = \text{radbas}(\|_i IW_{1,1} - p_N \| b_i1) \quad \ldots(4)$$

and the Competitive layer is given by

$$a2 = \text{compet}(LW_{2,1}a1) \quad \ldots(5)$$

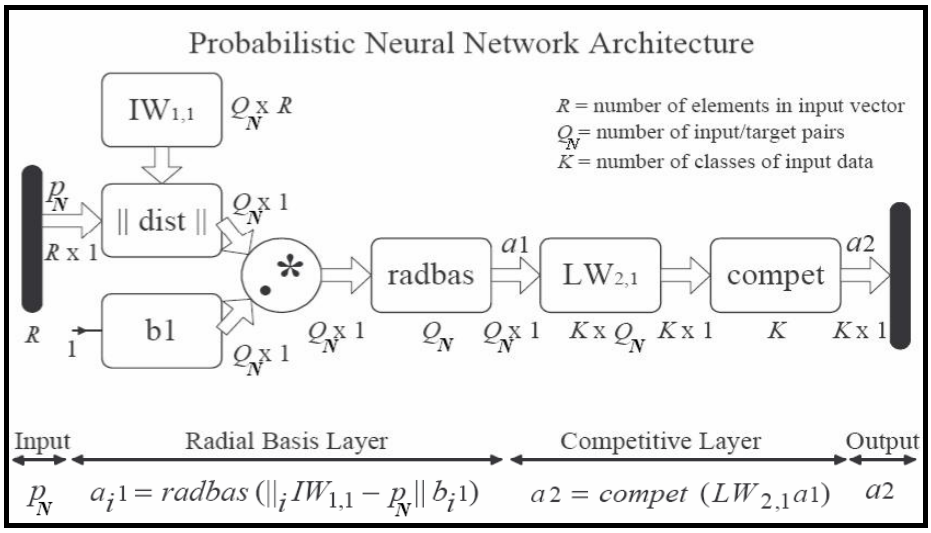

Figure 2: Architecture of the Probabilistic Neural Network, after [5].

where

$IW_{1,1}$ are the first layer input weights, set to be equal to the transpose of the matrix formed from the QN training vector pairs.

$LW_{2,1}$ are the second layer weights, set to be equal to the matrix of target vectors.

$\| . \|$ the Euclidian distance

i ith element of a1 & b1,the ith row of the IW1,1.

p_N the input feature vector

b_1 the bias for the Radial Basis layer, defined as [5]:

$$b_1 = \sqrt{-\ln(0.5)} / \sigma \quad \ldots(6)$$

a2 the binary output of the PNN second layer

The transfer function of the Competitive layer employs the winner-take-all rule. The biggest weighted sum of probabilities from the first layer is granted a '1', while the others receive zeros.

In the test phase (segmentation phase), the PNN classifier decides whether the input trial belongs to the first speaker or to the second speaker. In order to do this, the pre-trained

PNN is tested by the feature vectors extracted from the input speech. The degree of similarity of the input feature vector to either of the two speaker's model is estimated by computing their corresponding distances. For each input vector, a binary decision is made: output '1' means it belongs to the first speaker, while '0' is produced when the feature vector is more similar to the second speaker model.

Although the PNN network is used in the designing of this speaker segmentation system (which is known with its complexity and memory requirement) it has been found that the speaker segmentation system described above is capable of working in real time on common personal computers.

4.1 The training phase

For the purpose of segmentation, the two speakers participating in the conversation which will be segmented are known in prior. Therefore, in the training phase (which is performed before the segmentation or the testing phase) a PNN is trained by the two code books of the speakers participating in the conversation (refer to section 3). These two code books are pooled to get a (16*256) matrix which is used for training. A target matrix is built too, this target matrix contains one row (128 "1" followed by 128 "0"). As a result of the above, 210 PNN's are trained in this work to segment the 210 conversation used to test this segmentation system (refer to section 3).

4.2 The testing phase

In the testing phase, the PNN's that are obtained from the previous section are tested in three different categories, (male-male, female-female, male-female) each category has 70 conversations. To evaluate the system performance, the percentage of the false segmentations of the input segments (FS) to the total number of segments in the conversation under test (N) is used as a figure of merit referred to as (PFS).

$$PFS(\%)=FS/N * 100\% \quad \ldots.(7)$$

4.3 Segmentation results

For the three categories described above the segmentation results are demonstrated in Table 1. Three results are shown for each category. From the 70 segmentation trials performed for every category the maximum value of the percentage of the false segmentation (PFS) of a conversation, the minimum value of (PFS), and the average of all (PFS) for the 70 segmentation trials are found and they are shown in Table 1.

Table 1: The results for the segmentation system

CATEGORY	MAX. PFS(%)	MIN. PFS(%)	AVERAGE PFS(%)
Male-Male	31.8309	**11.4971**	22.8088
Female-Female	31.0738	16.6656	23.7910
Male-Female	**24.2105**	12.0422	**18.0936**

It is noticed form examining the segmentation results that there are some errors in the segmentation which are due to the silence periods between two word of the same speaker or a false segmentation of one or two segments during the speech of a certain speaker resulting in a "1" or two among a string of zeros or a "0" or two among a string of ones. Therefore, a new technique is added to the segmentation process to overcome this problem and take advantage

of these false segmentation to enhance the overall performance of the segmentation system as shown in the next subsection.

4.4 The enhancement of the segmentation system and its results

As mentioned earlier there are some errors in the segmentation which are due to the silence periods between two word of the same speaker or a false segmentation of one or two segments during the speech of a certain speaker. For an example the output of the neural network for an input phrase belonging to one of the speakers must be a stream of ones. But a pattern like the following could be found at the output of the neural network (11111101111110000111111110001111111). To enhance the proposed segmentation system and to overcome the problem of some of false segmented segments a sliding window of a finite length is moved along all the output of the neural network. If the number of 1's in this window is greater than the number of 0's a single "1" will represent this window and vice versa. The size of the window must be an odd number to avoid a situation where the number of 1's equal the number of 0's. Starting with the first two odd values after one, the values 3 and 5 are chosen for the size of this window. After moving the window over the output of the neural network, it is found by observing the results that there are still some errors that are similar to the first case (a "1" or two among a string of zero's or vice versa). Therefore, the sliding window is moved for a second time to overcome these errors and enhance the segmentation results. It must be noticed that each output of the neural network represent a 10ms frame of the input. Therefore, merging more than one output of the neural network in one (by using the technique described above) is not an infinite process. The reason behind this is that a speaker phrase can be as short as 200-250ms (for example the phrase "yes" or 'no"). Adding the silence periods after and before, a speaker can have a maximum length phrase of 350ms. Therefore, we stopped at a maximum window size of 150ms (the 3/5 or the 5/3 case window). The result of this enhancement step is illustrated in Tables 2,3,4.

Table 2: The results for the segmentation system for the male-male category

SIZE OF THE FIRST WINDOW / SIZE OF THE SECOND WINDOW	MAX. PFS(%)	MIN. PFS(%)	AVERAGE PFS(%)
0/0	31.8309	11.4971	22.8088
3/3	23.7504	4.6567	14.4969
3/5	23.9521	**2.8031**	**10.9496**
5/3	**21.8584**	2.9509	11.4644

Table 3: The results for the segmentation system for the female-female category

SIZE OF THE FIRST WINDOW / SIZE OF THE SECOND WINDOW	MAX. PFS(%)	MIN. PFS(%)	AVERAGE PFS(%)
0/0	31.0738	16.6656	23.7910
3/3	22.7051	9.9061	15.1781
3/5	20.0984	**2.6208**	**11.3113**
5/3	**20.0045**	3.4088	11.5922

Table 4: The results for the segmentation system for the male-female category

SIZE OF THE FIRST WINDOW / SIZE OF THE SECOND WINDOW	MAX. PFS(%)	MIN. PFS(%)	AVERAGE PFS(%)
0/0	24.2105	12.0422	18.0936
3/3	15.3839	3.1090	9.7224
3/5	13.2927	1.8487	**6.1389**
5/3	**13.1810**	**0.6825**	6.2379

5. Conclusions

The performance of a probabilistic neural network based segmentation system has been examined and enhanced in this paper. The system has been evaluated on the TIMIT database. The system has been examined in three different categories. It is found that for the situation where the two speaker participating in a conversation have different sex the system had its best performance with a minimum average percentage of false segmentation (PFS) equal to 18.0936%. To enhance the performance of the system a sliding window has been moved along all the output of the neural network twice. This step enhanced the performance of the segmentation system by reducing the value of PFS by approximately 66% for the different sex speaker system and by approximately 52% for the same sex speaker system. In addition it was found that a sliding window with a size of 3 for the first time and 5 for the second time gave the best enhancement for most of the results. Noting that the worst case (female-female) PFS is 11.3%. Comparing that with the best result of [1] which is (Miss Detection Rate) MDR =18.8% and (False Alarm Rate) FAR=21.8%. Where

MDR=MD/GT

FAR=FA/(GT+FA)

MD the number of miss detections

GT the actual number of speaker turns, i,e. ground truth

FA false alarms

PFS ≈ MDR + FAR.

6.References

[1] M. Kotti, L.G.P.M. Martins, E. Benetos, J.S. Cardoso, C. Kotropoulos, "Automatic speaker segmentation using multiple features and distance measures: a comparison of three approaches" ICME 2006 - IEEE 2006 International Conference on Multimedia & Expo, Toronto, Canada, July, 2006, pp. 1101-1104.
[2] M. K. Omar, U. Chaudhari, and G. N. Ramaswamy "Blind Change Detection for Audio Segmentation", Proc. of ICASSP-05, Philadelphia, Pennsylvania, March 2005, pp. 501-504.
[3] D. Wang, L. Lu, H.J. Zhang, "Speech segmentation without speech recognition", Proceedings of the 2003 IEEE International Conference on Acoustics, Speech and Signal Processing, 2003, pp. 468–471.
[4] M. Sharma, R. Mammone, "Automatic speech segmentation using neural tree networks" Proceedings of the 1995 IEEE Workshop on Neural Networks for Signal Processing, 1995, pp. 282-290.
[5] F. Gorunescu, "Benchmarking Probabilistic Neural Network Algorithms", International Conference on Artificial Intelligence and Digital Communication, Research Center for Artificial Intelligence, (2006).
[6] Kinnunen T., "Spectral Features for Automatic Text-Independent Speaker Recognition", Licentiate's Thesis, University of Joensuu, Finland (2003).
[7] Douglas A ., Thomas F., Robert B., "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing, Vol. 10, No. 1-3, pp. 19-51, (2000).
[8] Adami A., Hermansky H., "Segmentation of Speech for Speaker and Language Recognition Conditions", In Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003), pp. 841–844, (Geneva, Switzerland, 2003).
[9] Kinnunen T., "Optimizing Spectral Feature Based Text-Independent Speaker Recognition", Ph.D. thesis, University of Joensuu, Finland, (2005).
[10] Kinnunen T., "Spectral Features for Automatic Text-Independent Speaker Recognition", Licentiate's Thesis, University of Joensuu, Finland (2003).
[11] Garofolo J., Lamel L., Fisher W., "Darpa TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM Manual", National Institute of Standards and Technology (NIST), (1993).
[12] Ganchev T., Fakotakis N., Kokkinakis G., "One-speaker Detection – Limited Data: The WCL-1 System", NIST 2003 Speaker Recognition Workshop, College Park, MD, USA, (June24-25 2003).
[13] Ganchev T., Fakotakis N., Kokkinakis G., "Impostor Modeling Techniques for Speaker Verification Based on Probabilistic Neural Networks" Signal Processing, Pattern Recognition, and Applications (SPPRA 2003), Rhodes, Greece, (6/30/2003 - 7/2/2003).
[14] Ganchev T., Fakotakis N., Kokkinakis G., "Text Independent Speaker Verification Based on Probabilistic Neural Networks", In Proceedings of the Acoustics 2002,Patras, Greece, pp.159-166(2002).
[15] Demuth H., Beale M., "Neural Network Toolbox User's Guide", Version 4,MATLAB CD-ROM documentation, MathWorks Inc, pp. 7.12-7.20,(July, 2002).

The work was carried out at the college of Engg. University of Mosul