Meeting Segmentation Using Two-Layer Cascaded Subband Filters

Manuel Giuliani, Tin Lay Nwe, Haizhou Li

Institute for Infocomm Research
Republic of Singapore
manuel@manuelgiuliani.de, tlnma@i2r.a-star.edu.sg, hli@i2r.a-star.edu.sg

Abstract. The extraction of information from recorded meetings is a very important yet challenging task. The problem lies in the inability of speech recognition systems to be directly applied onto meeting speech data, mainly because meeting participants speak concurrently and head-mounted microphones record more than just their wearers' utterances - crosstalk from his neighbours are inevitably recorded as well. As a result, a degree of preprocessing of these recordings is needed. The current work presents an approach to segment meetings into four audio classes: *Single speaker*, *crosstalk*, *single speaker plus crosstalk* and *silence*. For this purpose, we propose Two-Layer Cascaded Subband Filters, which spread according to the pitch and formant frequency scales. This filters are able to detect the presence or absence of pitch and formants in an audio signal. In addition, the filters can determine how many numbers of pitches and formants are present in an audio signal based on the output subband energies. Experiments conducted on the ICSI meeting corpus, show that although an overall recognition rate of up to 57% was achieved, rates for crosstalk and silence classes are as high as 80%. This indicates the positive effect and potential of this subband feature in meeting segmentation tasks.

1 Introduction

Meetings are an important part of everyday work life. Many spend more time in meetings, where important goals and new strategies are discussed and determined, than on their desks. It is therefore desirable to extract the contents of a meeting and conserve them for future work or for purposes of proof. Automatic speech recognition (ASR), for example, seems to be a good tool to extract at least the textual content of a meeting. Unfortunately the recognition of speech in recorded meetings is a difficult task. Meeting participants speak naturally (i.e. use natural language), interrupt each other, talk at the same time and also use ungrammatical or incomplete sentences. In turn, these meeting conditions and norms, negatively affect ASR recognition rate. That is why some form of preprocessing of the recorded meetings is needed, among other things, to determine how many persons had been speaking at any one time in the meeting.

There have been some attempts at preprocessing of meetings. Dielmann and Renals [1] tried to segment meetings automatically into a set of social actions

such as *monologue*, *discussion* and *presentation*. For that purpose, they combined prosodical, lexical and speaker activity features to train and test a dynamic Bayesian Network model. With the so-called speaker activity feature, one can estimate which direction of the meeting room the speech recorded at a time is coming from. Therefore a microphone array was used to simulate a steerable directional microphone. Their experiments achieved a recognition rate of 92.9% and were conducted on the M4 corpus, which had been recorded at the IDIAP Research Institute. The M4 corpus contains 53 short meetings, recorded using lapel microphones for each meeting participant, and an eight element circular microphone array. However the lexical features that were used were based on human-generated word-level transcription of the meetings, entailing the employment of significant manual effort and that the segmentation process cannot be done automatically.

Wrigley et al. [2] segmented meetings into four different audio classes, namely *single speaker* (S), *crosstalk* (C), *speaker plus crosstalk* (SC) and *silence* (SIL). Crosstalk occurs when the lapel microphones or head-mounted microphones of meeting participants record not only their wearers' utterances, but also spoken comments from their neighbours. To discriminate the four different classes, Wrigley et al. analyzed several features on their efficiency for the task. Besides the classical speech processing features like MFCCs, Energy and Zero Crossing Rate, they also tested other features which had been proven to work well in similar tasks. These features include the following: Kurtosis, Fundamentalness, Spectral Autocorrelation Peak-Valley Ratio, Pitch Prediction, features derived from genetic programming and cross-channel correlation. After feature evaluation, they presented a system which consisted of a multistream ergodic Hidden Markov Model (eHMM) and a rule-based post processor to test the feature sets they had first found. They reported high average recognition rates of 76.5% for the speaker alone class, and 94.1% for the crosstalk alone class, but very low recognition rates for single speaker plus crosstalk and silence.

In the current work, we implement Two-Layer Cascaded Subband Filters (TLCSF) for meeting segmentation. The filters are able to extract the information of number of speakers based on the pitch and formant information. We combine this feature with other features which had been reported in [2] to classify the audio classes S, C, SC and SIL with higher accuracy. We trained Gaussian Mixture Models (GMM) for the four classes and linked them to an ergodic HMM. Experiment results show that our recognition rates are significantly higher for the classes SC and SIL.

The remaining of this paper is organized as follows: Section 2 describes the International Computer Science Institute (ICSI) meeting corpus which was used in our experiments. Following that is a presentation of our ergodic Hidden Markov Model in section 3, and an explanation of the acoustic parameters used in section 4. With the model and features in place, a range of experiments were conducted and are presented and discussed in section 5. Finally section 6 contains a conclusion of the current work as well as an outline of a few possibilities for future improvements.

2 Corpus

The ICSI Meeting Corpus consists of 75 meetings, which were recorded during the years 2000 - 2002 at the International Computer Science Institute (ICSI) in Berkeley, California. The meetings were not restricted by any guidelines, that means the recording sessions were held during normal meetings, which would have been conducted regardless of the recordings. In these recording sessions, every meeting participant wore either a head-mounted or a lapel microphone. At the same time, the meeting was recorded by six table microphones of different qualities. The meeting lengths range between 17 and 103 minutes and the corpus contains 72 hours of recorded speech in total. The data were collected at a 48 kHZ sample-rate, which was downsampled to 16 kHz. The audio files for each meeting are provided as separate time-synchronous recordings for each channel, encoded as 16-bit linear wave files and saved in the NIST sphere format. For each meeting, a time-tagged word-level transcript is available, which also contains meta information about its meeting participants and the hardware used for the session. A full description of the corpus can be found in [3]. From the corpus we chose 30 meetings, of which the data from 11 meetings were used to train the ergodic Hidden Markov Model and the data from the remaining 19 meetings were used for testing purposes.

3 Model

The model used in this work is an ergodic Hidden Markov Model (eHMM), which is made up of four GMMs, one for each of the four classes - S, C, SC and SIL - that we want to detect. The term *ergodic* refers to the fact that all four states of the HMM are linked together, such that every state is reachable from any other state and by itself, as illustrated in Figure 1. A GMM is defined by

$$\sum_{i=1}^{G} p_i \Phi_i(X, \mu_i, \Sigma_i) \tag{1}$$

where X is the feature vector and G is the number of Gaussian densities Φ_i. Every Φ_i has a mean vector μ_i, a covariance matrix Σ_i and a mixing coefficient p_i.

Every GMM was trained with the expectation-maximization algorithm, as it is implemented in the Hidden Markov Toolkit (HTK). The training data was extracted from 11 meetings [1] of the corpus. For each of the four classes, one million feature vectors were chosen randomly from the data. The number of mixtures per GMM varied and were chosen according to the values mentioned in [2]. The number of mixtures for the classes S and SC was set to 20, and to 5 and 4 for the classes C and SIL respectively. After training, the four GMMs were linked with transitions, such that every GMM is reachable from any other

[1] Training data was taken from the following meetings: Bed006, Bed008, Bed010, Bed011, Bmr001, Bmr005, Bmr006, Bmr014, Bmr024, Bro007, and Bro012

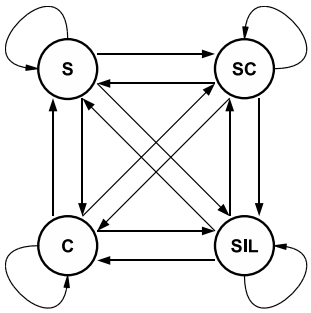

Fig. 1. *Ergodic Hidden Markov Model, comprising four GMMs*

GMM. The arcs between the GMMs were provided with transition probabilities, which were computed from the meeting transcripts.

4 Acoustic Parameters

The expressiveness of the acoustic parameters has direct impact on segmenting audio into different classes. In addition to short-time spectral information, we integrate pitch and formant information into our acoustic features. We propose Two-Layer Cascaded Subband Filters (TLCSF), which spread according to the pitch and formant frequency ranges. This filters are able to detect the presence or absence of pitch and formants in an audio segment. Furthermore, the filters can determine how many number of pitches or formants are present in an audio segment from the output subband energies. We transform these subband energies into cepstral coefficients for statistical modeling. The cepstral coefficients are used, because they have been proven to be robust in audio and speech recognition [7].

4.1 Acoustic Characteristics and Audio Classes

Before computing the features, we examine the significant characteristics possessed by each audio class. The signal strengths of the classes S (speaker alone) and SC (speaker plus crosstalk) are higher than those of class C (crosstalk alone) and class SIL (silence). In addition, the numbers of pitches and formants present in class S and class SC are different. The audio segment of class S has only one pitch or formant. However, the audio segment of class SC can present more than one pitch and formant. Furthermore, pitch and formant are not present in class SIL. Therefore, the acoustic features to identify these 4 audio classes (S, SC, C and SIL) should reflect the information on 1) signal strength, 2) the presence or absence of pitches and formants and 3) how many numbers of pitches and formants are present in an audio segment. To this end, we propose Two-Layer Cascaded Subband Filters to capture the above information from an audio signal.

4.2 Two-Layer Cascaded Subband Filters (TLCSF)

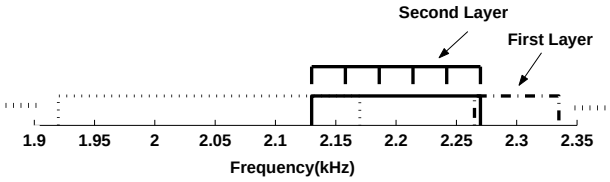

Fig. 2. *A bank of Two-Layer Cascaded Subband Filters.*

We propose Two-Layer Cascaded Subband Filters, shown in Figure 2, to capture the information of pitch, formant and signal strength. The filter has two cascaded layers. The first layer has overlapped rectangular filters. For each filter in the first layer, there are 5 non-overlapped rectangular filters of equal bandwidth in the second layer. The first filter of the first layer has a bandwidth spanned between 65Hz and 250Hz. This bandwidth covers the pitch of male and female in general [8]. This filter is able to determine the information on 1) presence or absence of pitch, and 2) number of pitches in an audio segment. Details on how the filter captures pitch information will be discussed in later paragraphs. Bandwidths of the following filters cover F1 (First formant), F2 (Second formant) and F3 (Third formant) of 15 selected English vowels [5]. Each of these filters determines the information on 1) presence or absence of formant, and 2) number of formants in an audio segment. To this end, we need to implement 1 filter for pitch and 45 filters for F1, F2 and F3 of each of the 15 vowels. Note that the formants of some vowels (example, First formants of vowels [aÈ] and [aØ]) overlap each other. Hence, we need to implement only one filter for these overlapped formants. Finally, we have 1 filter for the pitch (F0) and 40 filters for the formants (F1, F2 and F3). In Total, we implement 41 filters in the first layer. The center frequencies and bandwidths of all filters are listed in Table 1.

We have 41 filters in the first layer. For each filter of the first layer, we have 5 non-overlapped filters in the second layer. Hence, we have a total of 205 (41 x 5) filters in the second layer. The range of our subband filters is from 65 Hz to 3.2kHz.

The upper panels of Figure 3 (a), (b), (c) and (d) represent the signals of the four audio classes S, SC, C and SIL in the pitch frequency range (65Hz to 250Hz). As can be seen in the figures, the audio classes S and SC have the strongest signal strength of all four classes. In addition, only one pitch is present in audio class S and two pitches are present in the class SC, while no pitch is present in the class SIL. The Two-Layer Cascaded Subband Filter captures this information as follows.

The pitch information is captured by TLCSF for the four audio classes which are presented in Figures 3 (a), (b) (c) and (d). In each figure, the signal in the

No	Type	Vowel	CF(Hz)	BW (Hz)
1	F0	-	157.5	185
2	F1	[i]	300	72.5
3	F1	[u]	335	95
4	F1	[eÈ]	405	212.5
5	F1	[È]	435	120
6	F1	[ÿ]	445	150
7	F1	[oÈ]	455	260
8	F1	[Ø]	475	130
9	F1	[oØ]	495	170
10	F1	[aÈ], [a]	530	345
11	F1	[E]	575	150
12	F1	[O]	615	120
13	F1	[U]	620	80
14	F1	[Ï]	635	100
15	F1	[A]	700	130
16	F2	[oØ]	1000	270
17	F2	[O]	1015	150
18	F2	[u]	1075	460
19	F2	[E]	1085	360
20	F2	[Ø]	1140	180
21	F2	[A], [U]	1220	70

No	Type	Vowel	CF(Hz)	BW (Hz)
22	F2	[ÿ]	1290	100
23	F2	[oÈ]	1390	910
24	F2	[aÈ], [ÿ]	1540	765
25	F2	[Ï]	1575	295
26	F2	[E]	1605	240
27	F2	[È]	1700	300
28	F2	[eÈ]	1870	400
29	F2	[i]	2045	250
30	F3	[u]	2200	140
31	F3	[oØ]	2300	70
32	F3	[Ø]	2370	120
33	F3	[oÈ]	2425	195
34	F3	[Ï], [aØ]	2450	360
35	F3	[E]	2515	230
36	F3	[aÈ]	2525	250
37	F3	[U]	2550	140
38	F3	[eÈ]	2560	280
39	F3	[È],[O]	2585	170
40	F3	[A]	2600	160
41	F3	[i]	2960	400

Table 1. *Center Frequencies (CF) and Bandwidths (BW) of the 41 subbands in the first layer*

upper panel is passed through the TLCSF filters shown in the middle panel. Then, the output amplitudes of the five subband filters are computed and shown in the lower panel. As can be seen in the figures, the number of local maxima in the lower panel is the number of pitches present in the audio signal. Since TLCSF includes subbands for pitch and formant frequency ranges, these subbands work together to capture the pitch and formant information of the signal.

As mentioned above, formants of some vowels overlap each other. In Table 1, the filters with numbers 10, 21, 24, 34 and 39 are for 2 overlapped formants. Each of these filters covers the formants of two vowels. These filters can wrongly classify S as SC. The reason can be explained as follows: Let us assume, a speech segment of class S includes two vowels, [aÈ] and [aØ], which formants overlap in filter number 10. Two local maxima (two formants) can be present in the output of filter number 10 similar to the one shown in the lower panel of Figure 3(b). As we discussed above, if an audio segment has two local maxima, the classifier classifies the segment as 'SC'. Hence, we need to make sure that only one local maximum presents a segment of class 'S'. To this end, we choose an analysis frame length that covers the duration of only one vowel. For this reason, we choose the frame length of 60ms which is less than the average vowel duration [4].

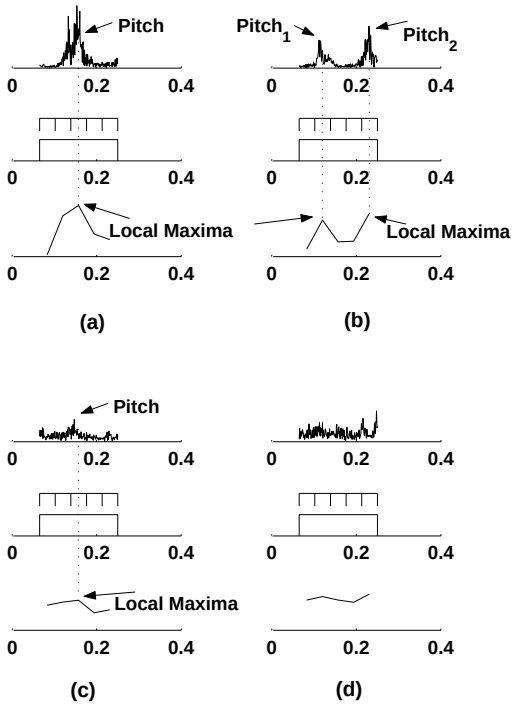

Fig. 3. *Capturing pitch information and TLCSF subband filtering: (a) Speaker alone - only one pitch and strong signal strength. (b) Speaker plus crosstalk - two pitches and strong signal strength. (c) Crosstalk alone - only one pitch and weak signal strength. (d) Silence - no pitch and weak signal strength. In each figure, the upper panel shows the signal. The middle panel presents the frequency response of the TLCSF subband filters. The lower panel demonstrates the output of the TLCSF subband filters. The filters capture the information on 1) presence or absence of pitch, and 2) number of pitches in the signal by detecting the local maxima.*

4.3 Computation of TLCSF coefficients

The speech signal was divided into frames of 60ms with 10ms overlapping. Each frame was multiplied by a Hamming window to minimize signal discontinuities at the end of each frame. Next, fast fourier transform (FFT) was applied, and following that, the audio frame was passed through a bank of cascaded subband filters and the log energy of each of 205 bands in the second layer was computed. Finally, a total of 40 Pitch and Formant Frequency Cepstral Coefficients (PFFCC) was computed from log energies using Discrete Cosine Transform [9] for each audio frame.

In Figure 4, example frames for the four classes S, SC, C and SIL and their corresponding feature vectors are illustrated. It can be seen clearly that the values of the feature vectors can be used to discriminate between the classes. For each class in Figure 4, two panels are shown. The top panels illustrate the

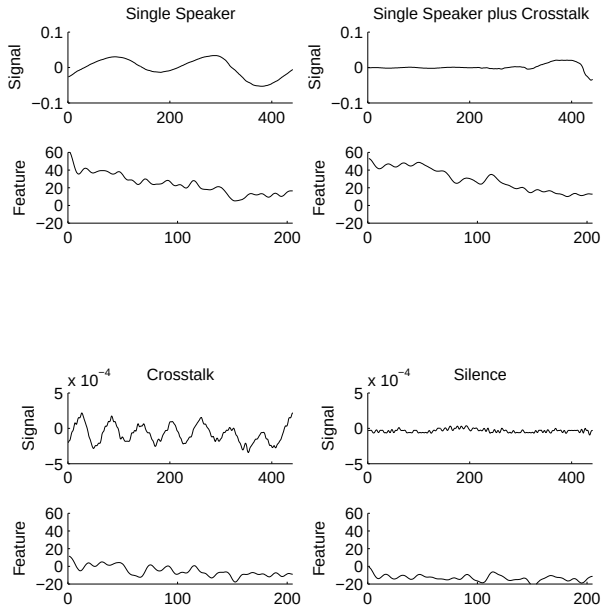

Fig. 4. *Illustration of signals and feature vectors for all four audio classes*

signal in time domain. Their corresponding subband feature vectors with 205 values are shown in the lower panel. Please note that the scales for the signals of S/SC and C/SIL differ for illustration purposes.

4.4 Features from Other Work

In addition to the PFFCC we introduced, seven other features were added. These have been shown to give good results in the meeting segmentation processes in [2]. Each of the following features was computed over a 16ms Hamming window with a frame-shift of 10ms.

- *Cross-channel Correlation* (CCC). The CCC is the maximum of the cross-channel correlation between a particular channel and any other channel. It was computed at any time of the signal. For each set of correlation values for any channel, the mean CCC, maximum normalized CCC and mean and maximum spherically normalized CCC was computed.
- *Kurtosis* (KU). Kurtosis is the fourth-order moment of a signal divided by the square of its second-order moment.
- *Log Energy* (LE).

5 Experiments

For the experiments, data from 19 of the ICSI meetings[2] were used. We tried combinations of different features to study the effects of the following parameters: Window length of the PFFCC feature, reduction of PFFCC features by dct, as well as a combination of the PFFCC features with the parameters *Cross-channel Correlation*, *Kurtosis* and *Log Energy*. As mentioned in Section 4.2, 60ms is a suitable window length for this task since this length is the average duration of a vowel. However, we would like to see the effect on the system performance using a shorter window length (example, 20ms). The reason is that a shorter window length could be a better choice to make sure that the audio segment includes only one vowel. Hence, we use window lengths of 20ms and 60ms to extract features. All these parameters led to six feature sets which are listed in Table 2. According to the window length and the number of features, the sets

Window Length	No. Total	No. PFFCC 40	No. PFFCC 205	CCC	KU	LE
20	41	•				•
20	46	•		•	•	•
20	211		•	•	•	•
60	41	•				•
60	46	•		•	•	•
60	211		•	•	•	•

Table 2. *Composition of feature sets*

are named 20-41, 20-46, 20-211, 60-41, 60-46 and 60-211.

The feature vectors from all the test meetings were extracted and labeled. Then recognition tests were made using the HTK. As we were interested to study only the effects of the feature combinations, no smoothing strategy was applied to the outgoing streams of recognition results.

In Table 3, we report the overall recognition results for two of the meetings (namely Bed015 and Bmr002) to show the performance of the different feature sets. It can be seen that the 41-dimensional feature sets, which contain the reduced PFFCC features and log energy, clearly outperform the remaining sets. But it should be noted that the low performance of the other feature sets may be due to a problem of normalization, which can be solved in future studies. In addition, we found that a 60ms window length performs better than a 20ms window. The reason for that is that a short window can not show a significant spectral difference between the different audio classes.

Figure 5 displays the recognition rate (line with circles) and the false positive rate (line with squares) for the single states of all meetings for the 60-41 feature

[2] The data of the following meetings were used: Bed015, Bed017, Bmr002, Bmr007, Bmr008, Bmr009, Bmr013, Bmr018, Bmr022, Bmr026, Bmr027, Bro008, Bro013, Bro014, Bro015, Bro017, Bro018, Bro023, Bro026.

Set	Bed015	Bmr022
20-41	**44.5%**	**46.0%**
20-46	33.9%	39.3%
20-211	22.7%	22.9%
60-41	**47.9%**	**52.7%**
60-46	28.6%	36.3%
60-211	19.0%	17.0%

Table 3. *Overall recognition results for meetings Bed015 and Bmr022*

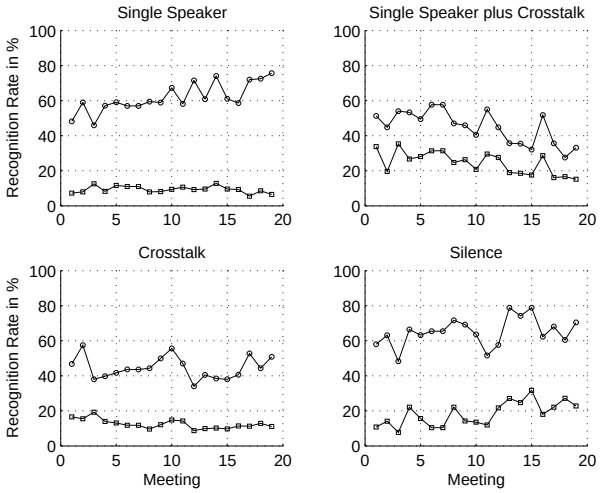

Fig. 5. *Recognition rates (upper line with circles) and false positive rates (lower line with squares) for all 19 test meetings using feature set 60-41. Each of the circles and squares stands for the recognition rate, and false positive rate accordingly, of one meeting.*

set, which performed best. The recognition rate denotes the percentage of correctly classified frames, while the false positive rate is specified as the proportion of negative instances that were erroneously reported as being positive.

These results also show that the recognition rate for the classes S and SIL are much higher than for the classes C and SC. Since our system aims to be used in the preprocessing of meetings for ASR systems, these results are very useful as they denote that a rather accurate detection of single speaker frames is possible and achievable. This indicates that the PFFCC feature is indeed suitable for the detection of several speakers and deserves further investigation. Our results can't be compared directly to the ones reported in [2], because on the one hand we used a slightly different set of the recorded meetings for training and testing. And on the other hand Wrigley et al. don't report their recognition results before applying a smoothing strategy.

6 Conclusions

In this paper, we presented a system for meeting segmentation, which segments recorded meetings into the four audio classes: *Single speaker*, *crosstalk*, *single speaker plus crosstalk* and *silence*. For that purpose we trained several ergodic Hidden Markov Models with different feature sets, which were made up of a feature that had been computed with two layers of subband-based filters, plus several other features that had been reported in other publications. Experiment results show that the performance of our system is effective for the single speaker and silence classes. Upcoming tasks to improve the recognition rate for the other classes can include normalization of the feature sets and implementing different models.

7 Acknowledgments

The authors like to thank Tomi Kinnunen for his help with the subband feature reduction.

References

[1] Alfred Dielmann and Steve Renals, "Multistream Dynamic Bayesian Network for Meeting Segmentation", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2006), May 14-19, 2006, Toulouse, France

[2] Stuart N. Wrigley, Guy J. Brown, Vincent Wan and Steve Renals, "Speech and Crosstalk Detection in Multichannel Audio", IEEE Transactions on Speech and Audio Processing, Vol. 13, No. 1, January 2005

[3] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke and Chuck Wooters, "The ICSI Meeting Corpus", Proc. ICASSP, 2003, pp. 364 - 367

[4] Xue Wang, Louis C.W. Pools and Louis F.M. ten Bosch, "Analysis of Context-Dependent Segmental Duration for Automatic Speech Recognition", International Conference on Spoken Language Processing (ICSLP), 1996, 1181-1184

[5] Dennis H. Klatt, "Software for a Cascade/Parallel Formant Synthesizer", J. Acoust. Soc. Am. 67, 971-995., 1980

[6] Haizhou Li and Tin Lay Nwe, "Vibrato-Motivated Acoustic Features for Singer Identification", Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, May 14-19, 2006, Toulouse, France.

[7] L. R. Rabiner, B. H. Juang, "Fundamentals of Speech Recognition", Prentice Hall, Englewood Cliffs, N.J, 1993

[8] G. Fant, "Speech Sounds and Features" Cambridge: MIT Press, MA, 1973

[9] C. Becchetti, L. P. Ricotti, "Speech Recognition Theory and C++ Implementation" New York: John Wiley & Sons, 1998