16

Soccer Event Retrieval Based on Speech Content: A Vietnamese Case Study

Vu Hai Quan
University of Science, VNU-HCM, Vietnam

1. Introduction

Video is a self-contained material which carries a large amount of rich information, far richer than text, audio or image. Researches (Amir et al., 2004), (Fleischman & Roy, 2008), (Fujii et al., 2006) have been conducted in the field of video retrieval amongst which content-based retrieval of video events is an emerging research topic. Figure 1 illustrates an ideal content-based video retrieval system which combines spoken words and imagery. Such ideal system would allow retrieval of relevant clips, scenes, and events based on queries which could include textual description, image, audio and/or video samples. Therefore, it involves automatic transcription of speech, multi-modal video and audio indexing, automatic learning of semantic concepts and their representation, advanced query interpretation and matching algorithms, imposing many new challenges to research.

There is no universal definition of video event, and the existing definitions can be classified into two types: one is being abnormal and the other is interesting to users (Babaguchi et al., 2002). In the first type of definition, an event may be either normal or abnormal. Generally speaking, only the abnormal event, which has more information than the normal one, is meaningful to the users. This event definition is suitable for the video analysis under restricted circumstance such as surveillance. The event definition of interesting to users is based on the users' description and domain prior knowledge (Sun & Yang, 2007). Suitable examples of this category are sport-video events such as ones in soccer and baseball. Several popular soccer events are shown in Fingure 2, including scoring, corner kick, yellow card and foul events.

Soccer video analysis plays an important role in both research and commerce. The basic idea of soccer events retrieval is to infer and retrieve the interesting events, and its goal is to make the results accord with human's visual perception as much as possible (Xu et al., 2001). Inference of events can be stemmed from either the semantic visual concepts or the spontaneous speech embedded in the videos. This chapter approaches soccer-video event retrieval in an audio aspect (i.e., the problem of spontaneous speech recognition). In this case, an event is defined as the spatiotemporal entity interesting to users, which is remarked by the announcer's spoken words. By exploiting spoken information of the video, soccer events are detected using an automatic speech recognition (ASR) system. However, as soccer videos vary in both speech quality and content, a canonical speech recognizer would not perform well without modifications and improvements. There are three main problems

induced by data diversity: noisy speech, foreign term interferences, and emotional variations in speech prosody.

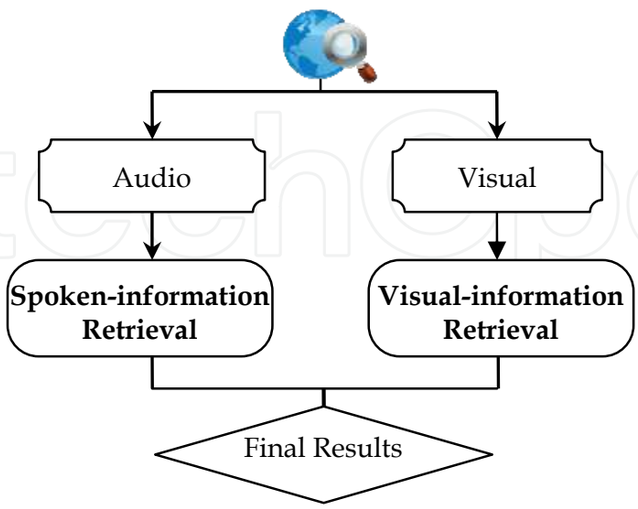

Fig. 1. A full-fledged content-based video retrieval system

To cope with these problems, a noise reduction scheme, a cross-lingual transliteration model, and an advanced acoustic modelling technique are proposed. In the remainder of this chapter, Section 2 gives a detailed specification of the retrieval system. Section 3 focuses on experimental evaluations. And finally, Section 4 concludes the discussions.

(a) Scoring (b) Corner Kick (c) Foul (d) Yellow Card

Fig. 2. Soccer events

2. The retrieval system

This section gives a detailed specification of the proposed retrieval system for soccer video database. Figure 3 illustrates four main parts comprising the system: a speech recognizer, a transliteration model, a noise suppressor, and a search engine. Each one plays an indispensable role in the whole system. The speech recognizer manages video transcriptions with the aid of transliteration model and the noise suppressor, while search engine deals with the tasks of indexing and retrieving transcribed text.

The following subsections will focus on each of the components respectively. As for the search engine, this system makes use of standard information retrieval techniques (e.g., indexing, matching, etc.). Therefore, it will not be covered in this chapter.

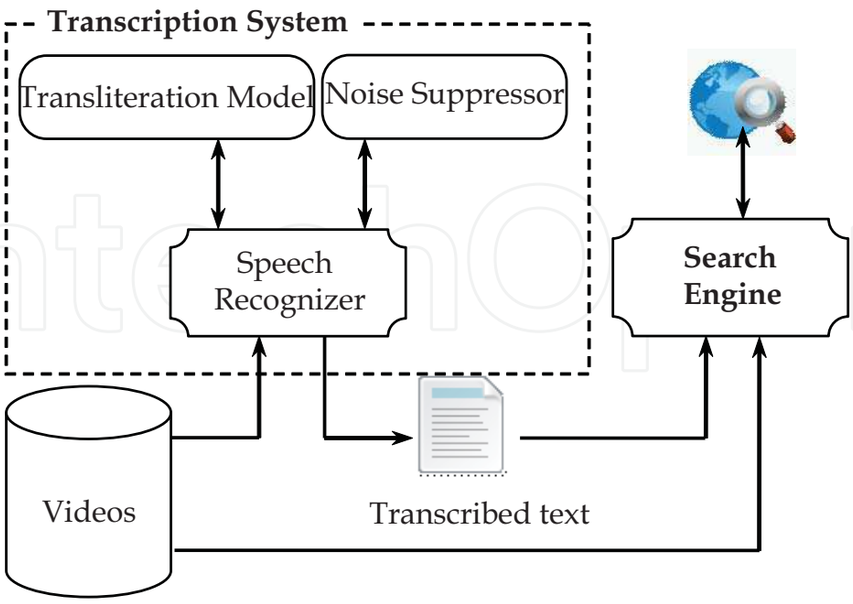

Fig. 3. The retrieval system

2.1 Vietnamese speech recognition

The speech recognition system that was reported in (Vu et al., 2006) is employed as the recognizer for the retrieval system. In this subsection, modifications in acoustic modelling and transcription process to the recognizer are discussed.

2.1.1 Advanced acoustic modeling

The acoustic modelling technique described in (Vu et al., 2006) is designed in the usual approach as for Chinese (Mori et al., 1997) in which each syllable is decomposed into initial (I) and final (F) parts (Figure 4a). While most of Vietnamese syllables consist of an Initial and a Final, some of them only have the Final. The initial part always corresponds to a consonant. The final part includes a main sound plus tone and an optional ending sound. This decomposition has two advantages. First, the number of monophones is relatively small (44 monophones). Second, by treating tone as a distinct phone, followed immediately after the main sound, the context-dependent model for tone can be built straightforwardly. It means that the recognition of tones was fully integrated in the system in just one recognition pass. However, distinct representations of tones have brought upon a disadvantage: the deficiency in modelling emotional variations of speech prosody. Since emotional prosody is expressed in the main tonal sound, separating tone from vowel would degrade the parameterization of tonal vowels.

To better model emotional prosodies, a modification to the acoustic model is proposed, in which tones are integrated into tonal vowels. This results in a new acoustic model consisting of 99 monophones including 27 phones for consonants, 12 phones for non-tonal vowels, and 60 phones for tonal vowels as shown in Figure 4b. Table 1 gives examples showing the differences between tone representations.

S → [I] F

F → V T [E]

I → b c ch d đ g gh gi h k kh l m n ng ngh nh p ph qu r s t tr th v x

V → a ă â e ê i o ô ơ u ư y

T → < / > < \ > < ? > < ~ > < . >

E → c ch ng nh m n p t

Fig. 4a. Separated tone modeling

S → [I] F

F → V [E]

I → b c ch d đ g gh gi h k kh l m n ng ngh nh p ph qu r s t tr th v x

V → a ă â e ê i o ô ơ u ư y
á ắ ấ é ế í ó ố ớ ú ứ ý
à ằ ầ è ề ì ò ồ ờ ù ừ ỳ
ả ẳ ẩ ẻ ể ỉ ỏ ổ ở ủ ử ỷ
ã ẵ ẫ ẽ ễ ĩ õ ỗ ỡ ũ ữ ỹ
ạ ặ ậ ẹ ệ ị ọ ộ ợ ụ ự ỵ

E → c ch ng nh m n p t

Fig. 4b. Integrated tone modeling

Word	Separated tone	Integrated tone
chào	ch a <\> o	ch à o
chao	ch a o	ch a o
chèo	ch e <\> o	ch è o

Table 1. Examples of tone representations

2.1.2 Video transcription

Speech in soccer videos is different from a typical speech training corpus in terms of quality and speaker-variations. This mismatch leads to serious degradation in system performance. In order to minimize errors, the soccer speech is put through a two-stage recognition process as shown in Figure 5. In the first stage, input speech along with its transcription, produced by the recognizer, are used to modify acoustic parameters. This is indeed the unsupervised

mode of acoustic adaptation. Gaussian components of the recognizer are adapted using the maximum likelihood linear regression (MLLR) technique. A global regression class is considered for adapting mean vectors with full transformation matrices. In the second stage, final transcriptions of input speech are generated by the adapted model.

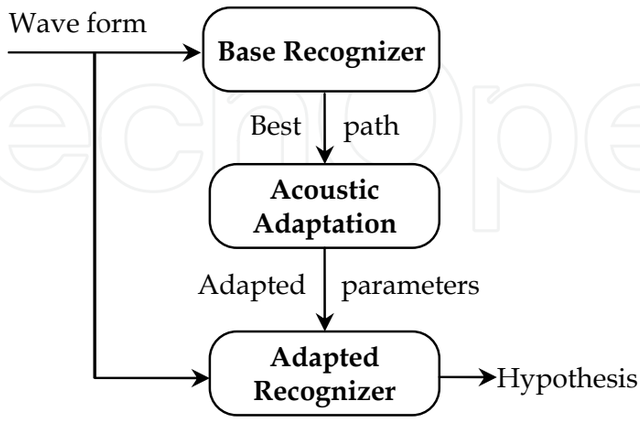

Fig. 5. Two-stage transcription process

2.2 Transliteration of foreign terms

The inability to deal with words in foreign languages causes recognition rates to drop drastically in ASR systems. A common solution to this problem is to look up a pronunciation dictionary. Despite its effectiveness, this approach has serious limitations: making a cross-lingual pronunciation dictionary of large size by hand is costly and required a lot of effort. Furthermore, the number of available entries is finite and therefore not flexible because speech recognition systems are expected to handle arbitrary words. Alternatively, data-driven approaches can be employed to overcome these limitations by learning samples and predicting unseen words. In the retrieval system, joint-sequence model (Bisani & Ney, 2008), a data-driven approach, is applied to transliterate foreign words into Vietnamese syllables.

The fundamental idea of joint-sequence model is based on the concept of graphone, a joint unit between graphemes and phonemes. In the assumption of joint- sequence model, each word and its pronunciation are generated by a common sequence of graphones, but the number of possible graphone sequences varies depending on the ways of segmentation. For instance, the word "David" and its pronunciation can be represented by one of the graphone sequences shown in Figure 6.

Graphone inventory can be estimated from training data using discounted EM algorithm (Bisani & Ney, 2008). The transliteration process searches for the most likely graphone sequence which matches the same spelling as given, and then projects it into phonemes. The resulting phonemes can then be assembled into Vietnamese syllables for speech recognition. It is worth noting that due to the co-segmentation characteristic of graphones, transliteration can be applied bidirectionally. It means that given a sequence of Vietnamese syllables, the corresponding foreign word can be obtained in the same way as presented.

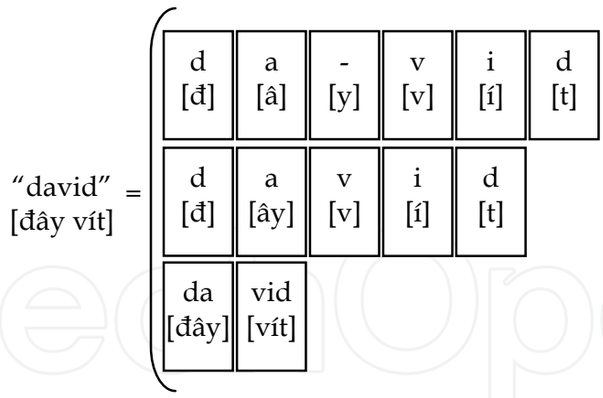

Fig. 6. Co-segmentations of the word "David" and its Vietnamese pronunciation

2.3 Noise reduction

Environmental variation has greatly affected the performance of ASR systems. A number of techniques have been proposed for dealing with environmental noise, especially additive noise which commonly plagues sport-domain speech. Additive noise is noise from external sound sources like wind or cheering that is relatively constant and can be modelled as a noise signal that is just added to the clean speech waveform to produce the noisy speech. One of the most popular methods for reducing the effect of additive noise is spectral subtraction (Katagiri et al., 1998). As depicted in Figure 7, the noise spectra S_n estimated during non-speech regions are subtracted from the noisy speech spectra S_y:

$$S_x = S_y - \alpha S_n \tag{1}$$

where α is the scaling factor for emphasis or de-emphasis of the noise spectra. Enhanced speech is then reconstructed based on the resulting spectra S_x.

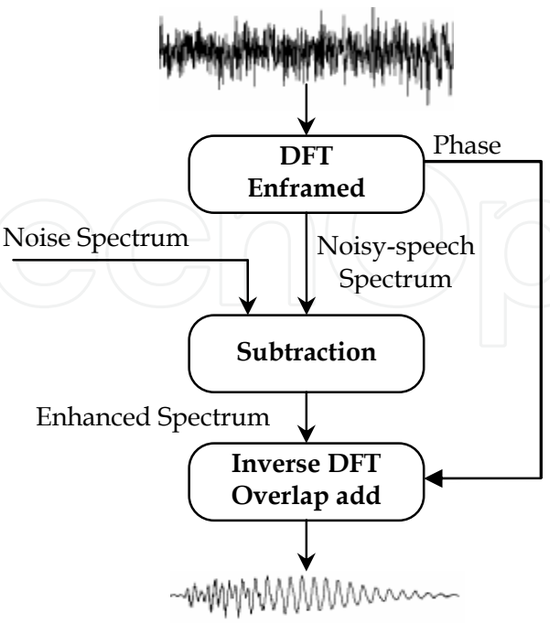

Fig. 7. Spectral subtraction

To minimize the word error rate induced by additive noise contained in soccer videos, magnitude spectrum subtraction is used to enhance speech quality of the videos. In addition, the smoothing technique, that was presented and proved in (Wojcicki et al., 2006) to be effective against the residual effect caused by spectral subtraction, is also employed.

3. Experiments

This section focuses on two main experiments: evaluations of the speech recognizer and the retrieval system. Both of them are conducted on the datasets described below.

3.1 Datasets

3.1.1 Speech and text corpora

The recognizers are trained with the speech corpus that was collected in 2005 from VOV – the national radio broadcaster (mostly in Hanoi and Saigon dialect), with a total duration of 20 hours. It was manually transcribed and segmented into sentences, which resulted in a total of 19496 sentences and a vocabulary size of 3174 words as shown in Table 2. All the speech was sampled at 16 kHz and 16 bits. They were further parameterized into 12 dimensional MFCC, energy, and their delta and acceleration (39 length front-end parameters).

Dialect	Duration	# Sentences
Hanoi	18 hours	17502
Saigon	2 hours	1994
Total	20 hours	19496

Table 2. The VOV speech corpus

Language models (bigram and trigram) for the recognizer are built using the 146M-word text corpus collected from newspaper text sources available on the Internet between 4/2008 – 10/2009. In addition, the text corpus (livescore – 2008) in soccer domain, consisting of 1M words, is also employed for language model adaptation.

3.1.2 Video database

For evaluation purposes, the AFF Suzuki-cup video database (2008) is demuxed into 14-hour speech channels. It is also manually transcribed and segmented into 11593 sentences, with a vocabulary size of 1810 words as shown in Table 3. The speech was sampled at multiple different rates, but was converted to an identical format of 16 kHz and 16 bits. This database will be served as the test-set for every experiment.

Dialect	Duration	# Sentence	# Foreign terms
Mixed	14 hours	11593	892

Table 3. The AFF video database

3.2 Evaluation metrics

Performance of an ASR system is typically measured in terms of word error rate (WER):

$$WER = \frac{S + I + D}{N} \tag{2}$$

where N is the total number of words in the test-set, and S, I, D are the total number of substitutions, insertions, and deletions respectively. This is indeed the edit distance between the automatically generated transcription and the reference one that was manually transcribed. This chapter makes use of word accuracy rate (WAR), which is defined as WAR = 1 – WER, to report performances of the recognizer.

In order to evaluate performances of the retrieval system, event-detection rates are measured in terms of recall and precision which are given by:

$$\text{Precision} = \frac{\#\ \text{correctly retrieved events}}{\#\ \text{retrieved events}} \tag{3}$$

$$\text{Recall} = \frac{\#\ \text{retrieved events}}{\#\ \text{relevant events in the database}} \tag{4}$$

3.3 Transcription evaluation

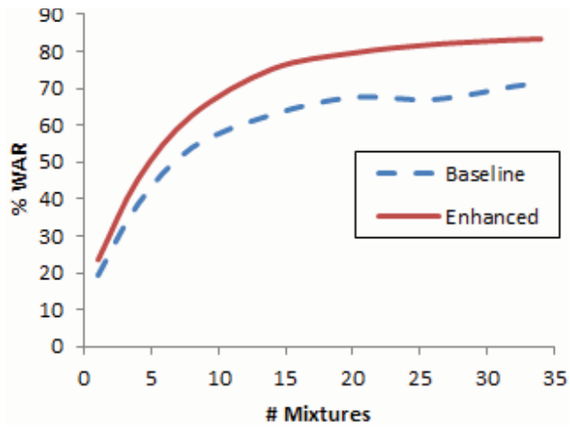

Fig. 8. Performances of the recognizers

In this experiment, the recognizer is evaluated on the task of soccer video transcription. To measure improvements obtained from the proposed methods of transliteration, acoustic modelling and noise reduction, the experiment is conducted in a comparative manner between the original recognizer (without any modifications presented in this chapter) and the modified one. Both are trained using the same corpora described in Subsection 3.1.1. Figure 8 plots performance functions of the two recognizers. As the number of Gaussian mixtures increases, the enhanced recognizer becomes dominant and an improvement of 11.9% can be seen in best case, where WAR reaches 83.3%.

3.4 Event detection evaluation

For evaluations of the retrieval system, Nutch[1] – an open source framework is deployed in role of the search engine. Several typical soccer events are selected as test cases, including: thẻ vàng (yellow card), thẻ đỏ (red card), phạt góc (corner kick), việt vị (offside), phạm lỗi (foul), ghi bàn/bàn thắng (scoring), Công Vinh (a Vietnamese player), Sukha (a Thai player). Table 4 reports their detection rates in the form of recalls along with the corresponding precisions.

[1]http://nutch.apache.org.

Event	# RlE*	# RtE**	# CRtE***	% Recall	% Precision
Yellow Card	32	20	17	62.50	85.00
Red Card	3	2	2	66.67	100.00
Corner Kick	56	38	32	67.86	84.21
Offside	48	34	30	70.83	88.24
Foul	121	65	59	53.72	90.77
Scoring	35	14	12	40.00	85.71
Cong Vinh	153	104	83	67.97	79.81
Su Kha	117	76	57	64.96	75.00
Average	**-**	**-**	**-**	**61.81 ± 9.56**	**86.09 ± 6.96**

*RlE: Relevant events
**RtE: Retrieved events
***CRtE: Correctly retrieved events

Table 4. Event detection rates

Most of the detection rates (i.e., recall) are above moderate while their precisions are pretty high. The average rates of 61.81% recall and 86.09% precision indicate a reasonable result for the proposed methods and their application in soccer event retrieval. This is indeed the single event detection mode in which each event is defined by a single keyword. Figure 9 gives examples of several single events and their false detections as well. Since events are remarked by the announcers' spoken works, errors in transcriptions will result in missing retrievals. And also, the context in which event-keywords are spoken will be responsible for the false detections. For instance, "scoring/goal" could be spoken in a regular comment (e.g., "vẫn chưa có bàn thắng/still no goal") rather than an authenticated scoring event.
Another way of retrieving soccer events is to combine several keywords together. These events will be denoted as "combined events." Figure 10 illustrates several combined events along with their false detections. Most of the false detections are caused by unexpected combinations between keywords in the results. For example, the combined query "Cong Vinh" & "yellow card" can be resulted in "a yellow card for player A for an unfair act with Cong Vinh" rather than the expected event "a yellow card for Cong Vinh." Someone may suggest enforcing phrase querying, but then again the phrase might not match the announcers' spoken phrase.

Event	# Retrieved events	# Correctly retrieved events	% Precision
"Cong Vinh" & "Yellow Card"	4	2	50.00
"Cong Vinh" & "Offside"	7	7	100.00
"Cong Vinh" & "Foul"	9	7	77.78
"Cong Vinh" & "Scoring"	10	2	20.00
Average	**-**	**-**	**61.95 ± 30.01**

Table 5. Performance of combined-event retrieval

Table 5 summarizes the performances of combined-event retrieval. Since the total number of retrieved events might exceed the number of relevant events, only precisions are reported.

(a) Scoring event
("**bàn thắng** mở tỉ số")

(b) False detection of scoring event ("vẫn chưa có **bàn thắng**")

(c) Foul event
("**phạm lỗi**")

(d) False detection of foul event
("không có **phạm lỗi**")

Fig. 9. False detections of single events.

3.5 Running-time evaluation

In this experiment, the retrieval system is evaluated in the manner of searching speed. Test cases/queries are generated randomly with respect to both single and combined events. Arbitrary phrase queries (with average syllable length of five) are also taken into account. Table 6 reports the average searching time for single keywords, combined keywords, and arbitrary phrases each with 200 different queries. All the tests were conducted in a standard server with a 16x3.02GHz processor and 32GB RAM.

Category	Searching time (seconds)
Single keyword	0.15
Combined keyword	0.24
Arbitrary phrase	0.31
Average	**0.23 ± 0.07**

Table 6. Average searching time

(a) "Cong Vinh" & "scoring" event ("**bàn thắng** của **Công Vinh**")

(b) False detection of "Cong Vinh" & "scoring" event ("chút xíu nữa là **Công Vinh** đã có **bàn thắng**")

(c) "Cong Vinh" & "yellow card" event ("một chiếc **thẻ vàng** cho **Công Vinh**")

(d) False detection of "Cong Vinh" & "yellow card" event ("với pha phạm lỗi với **Công Vinh** trước đó thì Alam Shah đã phải nhận **thẻ vàng**")

Fig. 10. False detections of combined events

A demo version of this system is available for testing at: www.ailab.hcmus.edu.vn

4. Conclusion

This chapter has presented a spoken information based approach for the retrieval of soccer video events – the first one to apply ASR in sport event retrieval. The entire retrieval system is centred on an automatic speech recognizer. To be applicable in the soccer domain, three modifications for the recognizer are proposed to resolve the problems of noisy speech, foreign term interferences, and prosody variations. Experiments on the video database give reasonable results for the proposed methods. In the near future, this system will be incorporated with the visual-information retrieval system to provide a flexible mechanism for the detection of semantic video events.

5. Acknowledgments

This work is part of the national key project no.KC01.16/06-10, supported by the Ministry of Science and Technology.

6. References

Amir, A., et al. 2004. A multi-modal system for the retrieval of semantic video events. *Computer Vision and Image Understanding*. 96, 2 (Nov. 2004), 216-236.

Babaguchi, N., Kawai, Y. and Kitahashi, T. 2002. Event based indexing of broadcasted sports video by intermodal collaboration. *IEEE Transactions on Multimedia*. 4, 1 (2002), 68-75.

Bisani, M., Ney, H. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*. 50, 5 (May 2008), 434-451.

Fleischman, M., Roy, D. 2008. Grounded language modeling for automatic speech recognition of sports video. In *Proceedings of ACL-08: HLT* (Columbus, OH, 2008). 121-129.

Fujii, A., Itou, K., and Ishikawa, T. 2006. LODEM: A system for on-demand video lectures. *Speech Communication*. 48, 5 (May 2006), 516-531.

Katagiri, E. S., Wan, E. A., and Nelson, A. T. 1998. Networks for speech enhancement. *Handbook of Neural Networks for Speech Processing* (1998).

Le, T., Nguyen, H., and Vu, Q. 2006. Progress in transcription of Vietnamese broadcast news. In *Proceedings of the International Conference on Communications and Electronics* (Hanoi, Vietnam, October 10 - 11, 2006). 300-304.

Mori, R. D., et al. 1997. *Spoken Dialogues with Computers*. Academic Press, San Diego, CA, USA.

Sun, X. H., Yang, J. Y. 2007. Inference and retrieval of soccer event. *Communication and Computer*. 4, 3 (Mar. 2007), 18-32.

Wojcicki, K. K., Shannon, B. J., and Paliwal, K. K. 2006. Spectral subtraction with variance reduced noise spectrum estimates. In *Proceedings of the 11th Australian International Conference on Speech Science & Technology* (Auckland, New Zealand, December 06 – 08, 2006). 76-81.

Xu, P., Xie, L. X., Chang, S. F., Divkaran, A., Vetro, A., and Sun, H. 2001. Algorithms and system for segmentation and structure analysis in soccer video. In *Proceedings of IEEE International Conference on Multimedia and Expo* (Tokyo, Japan, 2001). 576-579

Speech and Language Technologies
Edited by Prof. Ivo Ipsic

ISBN 978-953-307-322-4
Hard cover, 344 pages
Publisher InTech
Published online 21, June, 2011
Published in print edition June, 2011

This book addresses state-of-the-art systems and achievements in various topics in the research field of speech and language technologies. Book chapters are organized in different sections covering diverse problems, which have to be solved in speech recognition and language understanding systems. In the first section machine translation systems based on large parallel corpora using rule-based and statistical-based translation methods are presented. The third chapter presents work on real time two way speech-to-speech translation systems. In the second section two papers explore the use of speech technologies in language learning. The third section presents a work on language modeling used for speech recognition. The chapters in section Text-to-speech systems and emotional speech describe corpus-based speech synthesis and highlight the importance of speech prosody in speech recognition. In the fifth section the problem of speaker diarization is addressed. The last section presents various topics in speech technology applications like audio-visual speech recognition and lip reading systems.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen