MULTIPLE-FEATURE FUSION BASED ONSET DETECTION FOR SOLO SINGING VOICE

Chee Chuan Toh **Bingjun Zhang** **Ye Wang**
School of Computing
National University of Singapore
u0403701@nus.edu.sg, {bingjun, wangye}@comp.nus.edu.sg

ABSTRACT

Onset detection is a challenging problem in automatic singing transcription. In this paper, we address singing onset detection with three main contributions. First, we outline the nature of a singing voice and present a new singing onset detection approach based on supervised machine learning. In this approach, two Gaussian Mixture Models (GMMs) are used to classify audio features of onset frames and non-onset frames. Second, existing audio features are thoroughly evaluated for this approach to singing onset detection. Third, feature-level and decision-level fusion are employed to fuse different features for a higher level of performance. Evaluated on a recorded singing database, the proposed approach outperforms state-of-the-art onset detection algorithms significantly.

1. INTRODUCTION

Accurate monophonic singing voice transcription depends heavily on singing onset detection. Current onset detection algorithms applied on monophonic singing voice produced poor results [5][8]. In MIREX2007, the best result for onset detection of solo singing voice only managed an F-measure of 51.0% [2]. The singing voice is a pitched non-percussive (PNP) instrument [17], which is still a challenging category of instruments for onset detection. The case is further complicated by the nature of the singing voice, which is inherently inconsistent and prone to pitching and timing dynamics.

Current onset detection methods for PNP instruments, such as methods by spectral difference [7][10], phase deviation [6], pitch shift [9], and sub-band energy change [11][12] are targeted at detecting spectral changes based on certain rules. It has been observed on a spectrogram that in general, a musical note onset occurs at locations where there is a visible change in the spectrum, and within the duration of the note, the spectrum is relatively stable. A violin signal is shown as an example in Fig.1a.

However, for the case of singing voice, this observation may not always hold (Fig.1b). Unlike most other instruments, where there is usually a high level of timbre consistency in the duration of a note, the singing voice is capable of producing much more variations of formant structures (for articulation); sometimes the formant structure may even change within the duration of a single note. Pitch irregularities, pitch modulations, and inadvertent noise also upset stability of the spectrum.

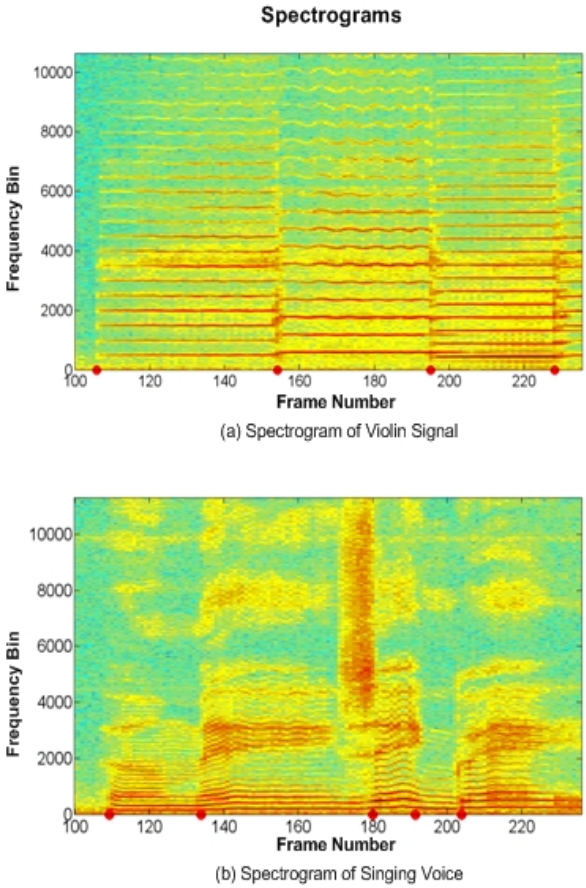

Figure 1: Comparison between spectrograms of violin and singing signal. Onsets are marked by red circles

Onset detection in the singing voice is much more complicated than in most other instruments. An onset detector that relies explicitly on a rule-based approach is bound to fail to capture the nuances of the singing onset. To address the difficulties in singing onset detection, we present a new onset detector based on supervised machine

learning, in order to capture the intricacies of singing note onsets, illustrated in Figure 2.

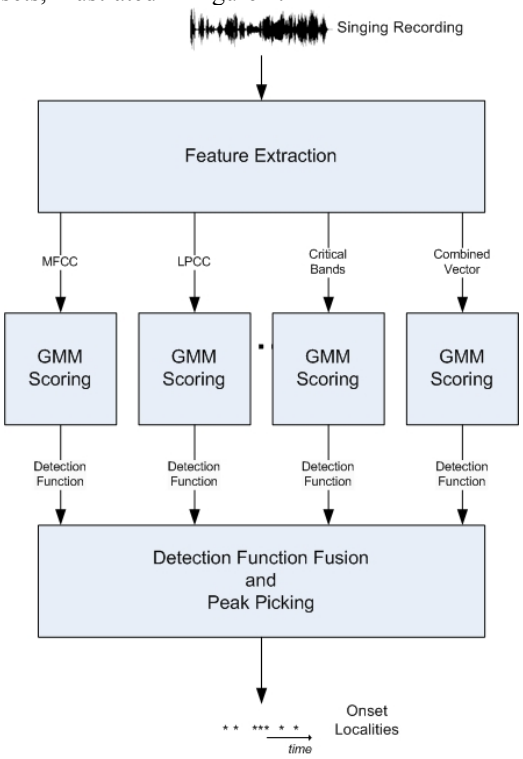

Figure 2: System diagram of the proposed system

Two GMMs are used to profile the distribution of audio features of onset frames and non-onset frames, respectively. We thoroughly evaluate existing audio features for their effectiveness in distinguishing onset and non-onset frames, including Mel-frequency Cepstral Coefficients (MFCCs), Linear Predictive Cepstrum Coefficients (LPCCs), Equal loudness phon values along critical bands. Feature concatenation fusion in feature-level and linear weighted sum fusion in decision-level are then employed to achieve a high level of onset detection accuracy. Evaluated using our singing database, our proposed approach to singing onset detection outperforms state-of-the-art methods, including methods based on phase change [6], pitch change [9], equal loudness [12], and inverse-correlation [7]. Re-implementations of these methods are used in the evaluation.

2. DEFINITION OF A SINGING ONSET

In a non-percussive continuous-pitch instrument, such as the melodic singing voice, there is no clear definition of an onset. A definition is especially difficult to establish when taking into account glissandos[1] and portamentos[2], which are vocal expressions involving pitch glides. Traditionally, an onset is the beginning of a new note event [5][8], characterized by a change in pitch, amplitude envelope, spectral shape, phase, or a combination of the mentioned factors [14].

However, in a singing voice, which is plagued by accidental glissandos, slurring (melisma[3]) and imperfect vocalization, the definition of an onset is a less clear-cut one. It is very often that a singer performs a glissando (more accurately, portamento in the singing voice) at the beginning or ending of a note, even when a music score does not stipulate so. When a musician is tasked to transcribe the singing performance, these portamentos are not transcribed into notes. However, slurred performances, which also involve pitch glides, are usually transcribed as actual notes, albeit with a legato denotation. There is no real physical difference between a portamento and a slur in singing performances. In both cases, the singing voice undergoes a continuous change in pitch. However, perceptually many glissandos are not considered as note onsets. This is because when listening to a singer's performance, a human subconsciously takes into account contextual cues like the rhythm, language and style of music.

The nature of singing onset is strongly related to the manner of articulation in speech. A singing voice has strong harmonic components producing numerous timbral differences, which are interpreted as different phonemes by the human ear. Singing onsets usually, though not always, occur during vowel-onsets.

Since there is no known existing definition for a singing onset, it is crucial to define one before even attempting to evaluate the effectiveness of a singing onset detector. In [5], an onset is defined as the start of the transient, but in a pitched singing voice (and many PNP instruments), there are numerous notes with no observable transient, particularly during slurs. A singing performance may also contain long stretches of unvoiced consonant segments, typically at the beginning or end of the note. These unvoiced segments are too varied and inconsistent to base the definition of a singing note onset on, and should simply be regarded as noise in the annotation process, not as part of a singing note. An onset should be marked at the end of a consonant, not before or during the consonant.

For practical purposes, it is intuitive and reasonable to define a singing onset as follows: The beginning of a new human-perceived note, taking into account contextual cues. This excludes erroneous portamentos during note beginnings, transitions and trail-offs, but includes pitch changes in slurring. The actual notes included in a slur

[1] Glissandos are glides from one pitch to another, by filling the slide with discrete intermediate pitches. In continuous-pitch instruments, glissando is often used interchangeably with the term portamento.

[2] Portamentos are continuous glides from one pitch to another, usually produced as an effect by continuous-pitch instruments.

[3] Melisma is the act of changing pitch on a single syllable of sung text.

may be subject to human interpretation, since it is possible that a singer reaches the correct pitch only very briefly during a slur. Slurred notes that occur too briefly for an average human auditory system to reliably detect should not be included. The precise location of an onset in a slur is also subjective, and allowances ought to be made for human bias.

3. FEATURE EXTRACTION

The effectiveness of any onset detection function ultimately depends on the audio features used in the system. It is necessary to employ features that provide useful information on whether or not a frame contains an onset.

Features capable of capturing timbral differences should provide a reliable measure for onset detection. Pitch and energy features could provide important information regarding onset detection as well, but are much less reliable. We have chosen to focus on features that provide information on the spectral shape, since a timbral change is the main characteristic of an onset.

3.1. Mel Frequency Cepstral Coefficients

MFCCs represent audio by a type of cepstral representation. In a mel-frequency cepstrum, the frequency bands are positioned on a mel-scale, which aims to approximate the human auditory system's response.

It is a common feature used in speech recognition, and can be derived by first taking the Fourier transform of a (windowed) signal, mapping the log-amplitudes of the resulting spectrum into the mel-scale, then performing DCT on the mel log-amplitudes. The amplitudes of the resultant cepstrum are the MFCCs. Since DCT holds most of the signal information in the lower bands, the higher coefficients can be truncated without excessive information loss. In our system, we extracted MFCC features with 81 mel-scale filter banks and 23 DCT coefficients.

The coefficients are concatenated with their first and second-order derivatives to form a feature vector of 69 dimensions. This improves the performance of the system, due to the correlation of the MFCCs and the derivatives at feature-level.

3.2. Linear Predictive Cepstrum Coefficients

LPCCs are LPC coefficients transformed into cepstra, and are widely used in speech recognition systems. An efficient method of obtaining the coefficients using Levinsin-Durbin recursion is covered in detail in [1]. Once we obtain the LPC coefficients, they can be transformed into cepstra by [4]:

$$c_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k} \quad (1)$$

where $\{c_1 \ldots c_N\}$ is the set of cepstral coefficients of order N. In our system, we used 22 cepstral coefficients, appended with 44 coefficients of the first and second order derivatives, similar to the feature-level fusion performed in the extraction of MFCCs.

3.3. Critical bands

Grouping frequency bins into critical bands is a psychoacoustically motivated principle. In [11], 36 bands of triangular-response bandwidth were used along the critical bands, and power in each band was used to derive a detection function. Detection functions based on equal-loudness changes in Equivalent Rectangular Bandwidth (ERB) bands have also been used in the past [8]:

$$\mathrm{E} = \lfloor 21.4 \log_{10} (4.37\,\mathrm{F} + 1) \rfloor \quad (2)$$

where E is the ERB band number, F is the bin frequency in kHz. By grouping frequency bins into ERB bands, we greatly reduce the dimensionality of the spectra. In [8], the power in individual ERB bands were mapped into phon values using equal-loudness contours [3]. In our system, we replicate the method introduced in [8], but only 36 ERB bands are used (includes frequencies of up to 11kHz) for audio files sampled at 22.05kHz.

3.4. Other Features Combined

Pitch-stability, zero-crossing rate and signal periodicity are all simple features which may contain information about note onsets. These features are commonly used, and can be concatenated into a combined feature vector. This feature-level data fusion is sensible for features which are characteristic of onset frames, especially if the features are of low dimensionality. Features which are of no value to the onset detection problem should not be included, as they corrupt the feature space and hence degrade the system performance. In our system, we used pitch stability, zero-crossing rate and signal periodicity and their derivatives as the combined vector.

4. GMM TRAINING AND SCORING

GMM-based modeling of speech signals has been the principal approach for speech and speaker recognition systems in recent years. It has been successfully implemented in [16] for speaker identification with very good results. Like speaker identification, onset detection can be modeled as a classification problem. At each time step, we need to classify a frame of audio into the onset class or a non-onset class.

Onset detection based on probabilistic models has increasingly been the preferred approach in recent audio research. Of particular interest are systems that employ machine learning algorithms that produced promising results, like Lacoste and Eck's FNN system [13]. Inspired by the success in speech processing systems, and in view of the similarity between the two problems, we employed a supervised machine learning algorithm using GMM classifiers.

4.1. GMM-based supervised machine learning

For each feature type, we model the probability of onset and probability of non-onset as random variables drawn from a Gaussian probability distribution of the feature vectors. Using a GMM to model the probability of onset random variable, we have:

$$P(x_n \mid \lambda_{onset}) = \sum_{i=1}^{M} w_i p_i(x_n) \quad (3)$$

where $P(x_n \mid \lambda_{onset})$ is the probability that feature vector x_n belongs to the onset class; w_i gives the weight of each of the M mixtures, and for each mixture:

$$p_m(x_n) = \sqrt{\frac{1}{(2\pi)^D \mid \Sigma_i \mid}} \, e^{-\frac{1}{2}(x_n-\mu_i)'(\Sigma_i)^{-1}(x_n-\mu_i)} \quad (4)$$

The parameters μ_i and Σ_i are the mean vector of dimensionality D, and the DxD covariance matrix, respectively. The parameters of the GMM can then be denoted $\lambda_{onset} = \{w_i, \mu_i, \Sigma_i\}$ where $i=1...M$. In our system, diagonal covariance matrices are used to parameterize the GMM, and 64 mixtures are chosen based on experiments. To train the onset GMM, the onset class training features are selected as the feature vectors of frames containing a hand-marked onset plus 3 frames at either side of each onset frame, which includes a total of 560 onsets (3920 frames).

We train this GMM by using the Expectation-Maximization (EM) algorithm [15] to profile the distribution of the onset-class feature vectors. We then proceed to construct another GMM, $\lambda_{non-onset}$, from the remaining feature vectors (21060 frames) for the non-onset class using (3) and (4).

4.2. Derivation of a Detection Function

With our trained GMMs (onset and non-onset classes), the probabilities $P(x_n \mid \lambda_{onset})$ and $P(x_n \mid \lambda_{non-onset})$ of each new feature vector can then be obtained. That is, for each new feature vector in time, we measure the likelihood that they belong to the onset and non-onset class. A viable detection function for our system will then be:

$$df(t) = P(x_t \mid \lambda_{onset}) - P(x_t \mid \lambda_{non-onset}) + 1 \quad (5)$$

where x_t is a feature vector of the audio signal obtained at time t. The detection function is then normalized to a range of [0…1].

5. DETECTION FUNCTION FUSION AND PEAK-PICKING

As aforementioned, for each feature type, we train 2 GMMs (for onset-class and non-onset-class features). Since we could have several feature types, we train a pair of GMMs for each feature type and derive a detection function for them using (5). Thereafter, we linearly weigh each of the detection function, and compute a sum of the individual weighted detection functions to produce a combined detection function:

$$df(t) = \sum_{i=1}^{F} w_i df_i(t) \quad (6)$$

F is the number of feature types, and w_i gives the weight of each of the detection functions.

It is also possible to use only a pair of GMMs, by simply concatenating the dimensions of each of the individual feature vector into a single feature vector at the feature-level, but solely relying on such an approach will increase the dimensionality of the feature space considerably, and it will require exponentially more training data in order to fully train the GMMs, according to the curse of dimensionality[1]. It is therefore prudent to keep each of our feature space limited to a feature type and its derivatives, possibly except for simple features of very low dimensions.

Once we obtain the detection function described by (6), we apply a de facto standard median-filter peak-picking algorithm, used in both [5] and [8] to evaluate the detection functions. The output of the peak-picking process is a series of onset locations denoting the time points at which an onset has occurred.

6. EXPERIMENTAL RESULTS

6.1. Database Description

Our database consists of 18 singing recordings of pop songs, from 4 singers (2 male, 2 female) of varied singing styles. The pieces were annotated by hand for onsets, and contained a total of 1127 onsets. Approximately half

[1] Curse of dimensionality is a term coined by Richard Bellman (1920-1984) describing the exponential relationship between volume and dimensionality.

(560) of the onsets were used for training GMMs, and the other half (567) were used to evaluate the system.

Each recording was annotated and cross checked by two amateur musicians. Onsets were first identified by ear, and then marked at positions where the wave form was first observed to follow a periodic structure. Consonant noise and unintentional portamentos / glissandos do not constitute new onsets, as explained in Section 2. For legato, an onset was marked at the point where a pitch change was perceived to begin.

6.2. Evaluation of Individual Detection Functions

For evaluation of the individual onset detection functions, we extract all the features mentioned in Section 3 and score each feature type based on a pair of trained GMMs (onset and non-onset GMMs). From the detection function produced by (5) for each feature type, we apply the median-filter peak-picking process across different parameters. We evaluate the usefulness of the feature by the metrics of precision, recall and F-measure, based on an onset tolerance of 50ms. Based on an onset tolerance of 50ms. These evaluation conditions are identical to those in the annual MIREX Audio Onset Detection contest.

The onset detection results of individual features are shown in Figure 3. The best F-measure among the set of peak-picking parameters is selected for each feature. From our experiments, MFCC is revealed to be the best audio feature, producing the best results of 80.3% precision, 77.8% recall and 79.0% F-measure. LPCC is next best, with 72.9% precision, 77.7% recall, and F-measure 75.3%. ERB-bands produce 78.2% precision, 69.5% recall, and 73.6% F-measure. The combined vector produces 60.1% precision, 86.1% recall, and 70.8% F-measure.

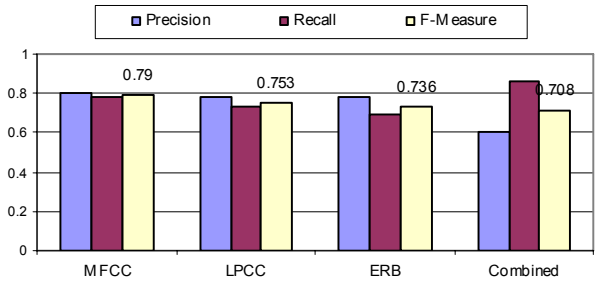

Figure 3: Evaluation results of individual detection function

6.3. Evaluation of Combined Detection Function

Based on the performance of all evaluated detection functions, we search the best linear weights based on extensive experiments and compute an overall detection function described by (6). The weight distribution is shown in Table 1.

Detection Function produced by:	Weight
MFCC	0.76
LPCC	0.12
ERB-bands	0.10
Combined vector	0.02

Table 1: Weight distribution for detection functions

Using the sum of linearly-weighted detection functions, we achieve the best performance of 86.5% precision, 83.9% recall, and 85.2% F-measure, which is superior to current state-of-the-art methods. As can be seen in Figure 4, most existing methods do not perform well on singing music. Equal loudness change based method produces an F-measure of 71.0% under the best set of parameters. Both the phase-based and pitch-based methods perform badly because the singing voice's pitch track is very unstable, and contains many noisy segments.

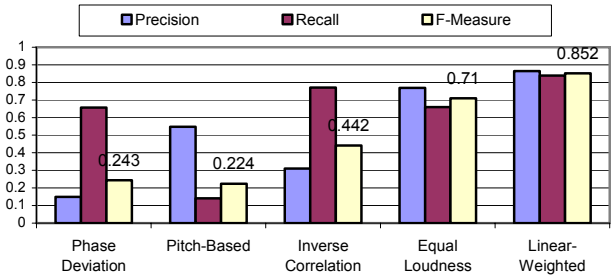

Figure 4: Results of linear-weighted sum detection function compared with state-of-the-art methods

The parameters for the system, especially those used in the GMM and individual features, were determined through extensive experimentation.

7. DISCUSSIONS AND FUTURE WORK

In our system, we have employed both feature-level and decision-level fusion. Feature-level fusion works well, but the higher dimensionality necessitates higher volume of training data. Our system was trained with only 560 onsets, due to the laborious process annotating pieces by hand. More training and testing data ought to be used to validate the system's performance.

We utilized only 4 pairs of GMMs: for MFCC, LPCC, ERB-bands, and a combined feature. In reality, any number of feature types can be used, and fusion can be

done at the feature-level (e.g. concatenation) and/or at decision-level (e.g. linear weighted sum). As a rough guide, more feature types and detection functions usually produce better onset detection results. This is provided the features extracted are representative of the onset classification problem, i.e. there should be a discernible difference in feature between onset and non-onset frames.

Weight-assignments for linear-weighted sum fusion are usually based on heuristics, and require lengthy experiments to optimize. Even though it generally works well, the simple linear weighting method can cause false peaks in detection functions to propagate into the combined detection function, making the overall detection function noisy. Other decision-level fusion techniques exist, and ought to be explored and tested.

We also look forward to expanding the system by incorporating note segmentation, as well as more functions in the post-processing section.

8. CONCLUSION

As shown by our experiments, the proposed supervised machine-learning approach based on GMM modeling produces higher accuracy for singing onset detection. Clearly, the system produces better results than state-of-the-art singing voice onset detection algorithms. Further improvements by decision-level fusion of the system boost the overall accuracy and completeness of the system. The singing onset detection problem cannot be considered solved by our system, but its potential is promising.

9. ACKNOWLEDGEMENT

This work was supported by Singaporean Ministry of Education grant with the Workfare Bonus Scheme number of R-252-000-267-112.

10. REFERENCES

[1] Sung-Won Park, *online DSP course,* http://www.engineer.tamuk.edu/SPark/course.htm

[2] MIREX 2007 Audio Onset Detection Results: Solo Singing Voice, http://www.music-ir.org/mirex/2007/index.php/Audio_Onset_Detection_Results:_Solo_Singing_Voice

[3] ISO Acoustics: Normal-equal loudness contours. Technical Report ISO226:2203, *International Organisation for Standardization*, 2003

[4] G. Antoniol, V. Rollo, G. Venturi "Linear Predictive Coding and Cepstrum coefficients for mining time variant information from software repositories" *International Workshop on Mining Software Repositories, ACM,* 2005

[5] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, M. Sandler "A Tutorial on Onset Detection in Music Signals" *IEEE Transactions on Speech and Audio Processing,* volume 13(5), September 2005

[6] J. Bello and M. Sandler "Phase-Based Onset Detection for Music Signals", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 49-52, April 2003.

[7] W. Boo, Y. Wang and A. Loscos "A violin music transcriber for personalized learning.", Proceedings of IEEE International Conference on Multimedia and Expo, pages 2081-2084, 2006

[8] N. Collins "A comparison of sound onset detection algorithms with emphasis on psycho-acoustically motivated detection functions", *Proceedings of AES118 Convention*, 2005

[9] N. Collins "Using a Pitch Detector for Onset Detection", Proceedings of 6th International Conference on Music Information Retrieval, 2005

[10] C. Duxbury, M. Sandler, M. Davies "A Hybrid Approach to Note Onset Detection", *Proceedings of the 5th Int. Conference on Digital Audio Effects (DAFx-02)*, Hamburg, Germany, 2002

[11] A. Klapuri, A. Eronen, J. Astola "Analysis of the Meter of Acoustic Musical Signals", *IEEE Transactions on Speech and Audio Processing,volume 14(1), pages 342-355,* January 2006

[12] A. Klapuri. "Sound onset detection by applying psychoacoustic knowledge." *In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 3089–3092, 1999.

[13] A. Lacoste and D. Eck "A supervised classification algorithm for note onset detection" *EURASIP Journal on Advances in Signal Processing*, August 2007

[14] A. Loscos "Spectral Processing of the Singing Voice", *Ph.D. Thesis submission to Pompeu Fabra University*, Barcelona, Spain, 2007

[15] R. Redner, H. Walker. "Mixture densities, maximum likelihood and the EM algorithm" *SIAM Review,* volume 26(2), 1984

[16] D. A. Reynolds and R. C. Rose "Robust text-independent speaker identification using Gaussian mixture speaker models", *IEEE Transactions on Speech and Audio Processing*, volume 3(1), pages 72-83, January 1995

[17] J. Sundberg, *The Science of the Singing Voice*, Northern Illinois University Press, 1990.