Multi-Stage Feature Normalization for Robust German Stressed/Unstressed Syllable Classification

Yuan-Fu Liao, Yan-Ting Chen and Jhen-Lun Huang

Institute of Computer and Communication Engineering, National Taipei University of Technology, Taiwan
yfliao@ntut.edu.tw, ted01261986@hotmail.com, lun77812@hotmail.com

Abstract

To develop a German computer assisted language learning (CALL) system for students whose mother's tongues are syllable- or mora-timed, a multi-stage feature normalization scheme which takes both word stress and sentence intonation patterns into consideration is proposed for German stressed/unstressed syllable classification. The main idea is to first apply Fujisaki model and band-pass filtering to pitch and energy contours, respectively, to remove the undesired sentence intonation component and sequentially normalize the extracted features in syllable- and supra-segment-level. Comparing with traditional Z-Score feature normalization baseline, the proposed method achieved lower classification error rate (27.04% vs. 31.34%) on "The Kiel Corpus of Read Speech, Vol. I" database. Besides, by integrating decision tree-based feature selection and long-span contextual prosodic cues, the system performance was further improved to 24.68%.

Index Terms: prosodic feature normalization, German stressed/unstressed syllable classification, Fujisaki model

1. Introduction

Since German is a stressed-timed language, putting the stress on the wrong syllable is more likely to make a word unintelligible than is mispronouncing one of its sounds. Especially, for students whose mother's tongues are syllable- or mora-timed, misplaced syllable stress is often their main problem to master German. Therefore, a German CALL program needs to be able to automatically identify and correct those pronunciation mistakes. To this aim, this study focuses on classifying German stressed/unstressed syllable and treats it as a strict two-class problem, applied only to individual syllables taken out of their word context.

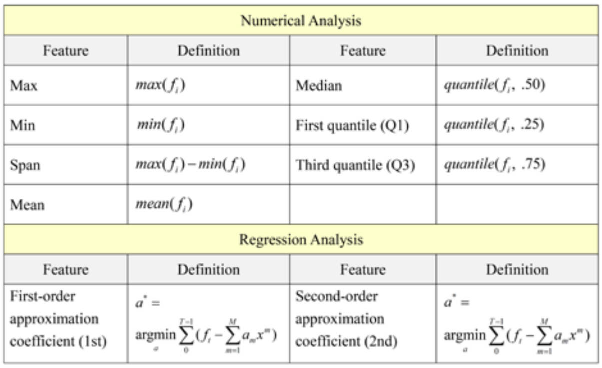
In the past, many prosodic features have been used to successfully identify stressed/unstressed syllable, because the stressed syllables usually have larger sound, longer time and change of pitch on pronunciation. For example, in [1-4], duration, fundamental frequency (F0) and energy contours of a syllable nucleus were first measured and then their derivations, such as max, min, mean, span, etc. (defined in Table 1) were extracted as the set of raw prosodic features.

It is worth noting that the extracted raw prosodic features are often affected by many factors not related to stress. So, it is essential to do some extra processing to remove those unwanted interferences. For example, in [5-10], feature normalization methods, such as mean subtraction and unity variance, i.e., Z-Score are usually applied in syllable- or word-level to generate a set of more robust prosodic features, i.e.,

$$x' = \frac{\left(x - mean(x)\right)}{\sqrt{var(x)}} \tag{1}$$

Here x and x' are the raw and normalized prosodic features (see Table 1), respectively.

Table 1. *The whole set of prosodic features extracted for stressed/unstressed syllable classification.*



Numerical Analysis			
Feature	Definition	Feature	Definition
Max	$max(f_i)$	Median	$quantile(f_i, .50)$
Min	$min(f_i)$	First quantile (Q1)	$quantile(f_i, .25)$
Span	$max(f_i) - min(f_i)$	Third quantile (Q3)	$quantile(f_i, .75)$
Mean	$mean(f_i)$		
Regression Analysis			
Feature	Definition	Feature	Definition
First-order approximation coefficient (1st)	$a^* = \underset{a}{\operatorname{argmin}} \sum_{0}^{T-1}(f_t - \sum_{m=1}^{M} a_m x^m)$	Second-order approximation coefficient (2nd)	$a^* = \underset{a}{\operatorname{argmin}} \sum_{0}^{T-1}(f_t - \sum_{m=1}^{M} a_m x^m)$

However, the extracted prosodic features are affected not only by the word stress but also the sentence intonation pattern. Therefore, it may be not enough to do only syllable- or word-level feature normalization.

So, in this paper, a multi-stage feature normalization scheme which takes both word stress and sentence intonation patterns into consideration is proposed. The main idea is to (1) first apply Fujisaki model [4] and band-pass filtering for pitch and energy contours, respectively, to remove the undesired sentence intonation (or phrase) component and extract prosodic features on the desired word stress (or accent) component; (2) then apply feature normalization in syllable-level, (3) compute long-span contextual prosodic cues (a kind of difference features) in supra-segment-level (or word-level) and (4) apply feature normalization again. The overall block diagram of the proposed approach features is shown as follows:

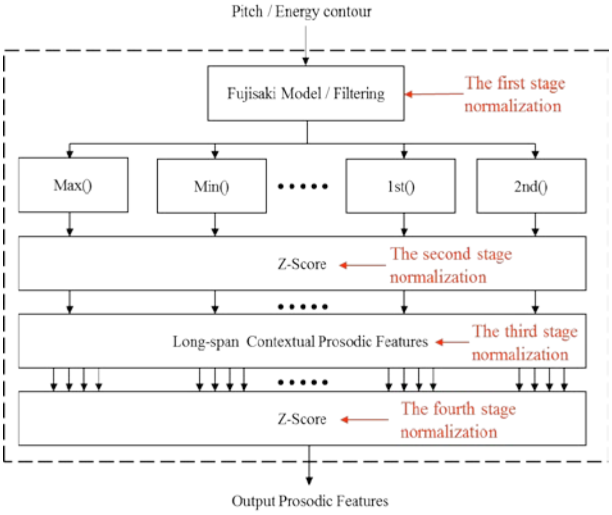

Pitch / Energy contour
Fujisaki Model / Filtering
The first stage normalization
Max()
Min()
•••••
1st()
2nd()
Z-Score
The second stage normalization
•••••
Long-span Contextual Prosodic Features
The third stage normalization
•••••
Z-Score
The fourth stage normalization
Output Prosodic Features

Figure 1: *The schematic diagram of the proposed multi-stage feature normalization approach for pitch and energy-related prosodic feature extraction.*

The rest of this paper is organized as follows. Section 2 describes the proposed multi-stage feature normalization approach. Section 3 reviews decision tree-based feature selection method. Section 4 reports the experimental results evaluated on Kiel German reading speech corpus [12]. Some conclusions are given in the last section.

2. Stressed/Unstressed Syllable Classifier

Fig. 2 shows the block diagram of the proposed stressed/unstressed syllable classifier as a module of a German CALL system. It includes (1) an aligner to find the positions of syllable nuclei, (2) a feature extraction and multi-stage feature normalization module (i.e. the block diagram in Fig. 1) to compute a set of robust prosodic features for each syllable nucleus and (3) a stressed/unstressed syllable recognizer to classify each syllable into stressed or unstressed one.

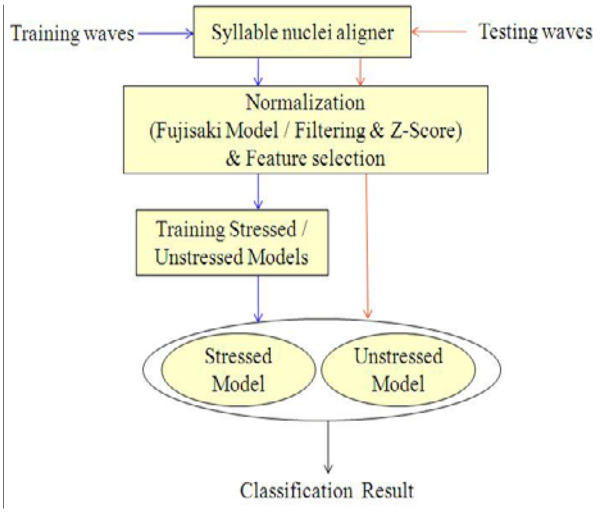


Figure 2: *The block diagram of the proposed stressed/unstressed syllable classifier.*

First of all, since the goal is to develop a stressed/unstressed syllable classification module for German CALL system, it is safe to assume that the aligner has prior knowledge of said prompts and their transcriptions to optimize pronunciation evaluation.

Secondary, in this study, pitch and energy contour related prosodic features are extracted using the procedure shown in Fig. 1. The only difference between pitch and energy contour processing is that for energy-related features, a specific band-pass filter is designed and used instead of the Fujisaki model. Therefore, in the feature extraction and normalization module (Fig. 1), Fujisaki model and band-pass filtering is first applied to pitch and energy contours, respectively, to remove the undesired sentence intonation component. Then a set of prosodic features as defined in Table 1 is extracted and normalized using syllable-level Z-score method. In this stage, there are 9 pitch-, 9 energy- and 1 duration-related features (in total 19 dimensions).

Thirdly, a set of 76-dimensional long-span contextual prosodic cues (difference features) are calculated in supra-segment-level from the set of 19 features and further normalized using Z-Score method.

In the following subsections, we briefly introduce (1) Fujisaki model analysis and (2) longer-span contextual prosodic cues extraction modules.

2.1. Fujisaki Model

The Fujisaki model, shown in Fig. 3, assumes that the F0 contour (in a logarithmic scale) is the superposition of three contributions: a pitch frequency baseline (Fb), a phrase component (Ap) and an accent component (Aa), obtained by filtering two input signals.

The first contribution (Fb) represents the pitch baseline of an utterance. The second contribution (Ap), which models the speaker declination and is characterized by a fast rise followed by a slower fall. The third contribution (Aa), which models smaller-scale prosodic variations, accounts for accent components.

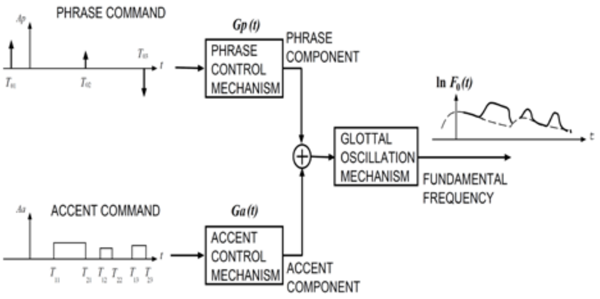


Figure 3: *Block diagram of the Fujisaki-model (from [11]).*

Fig. 4 shows a typical result of the Fujisaki model analysis for an input pitch contour of a German utterance. Here the pink dots and blue lines in the second panel represent the raw and smoothed pitch contours, respectively. On the other hand, green and orange lines in the third and fourth panel are the phrase (Ap) and accent (Aa) components, respectively. Among them, only (Aa) component is related to syllabic stress and will be used for feature extraction.

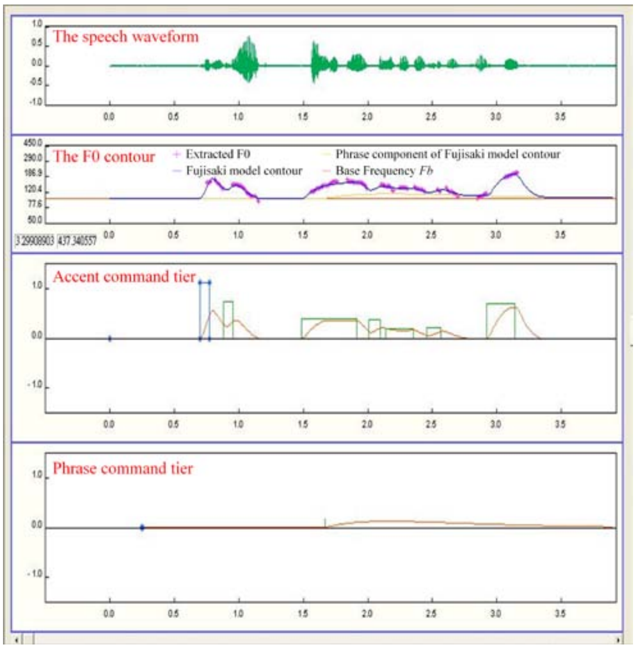


Figure 4: *A typical example of Fujisaki model decomposition for an input pitch contour of a German utterance.*

2.2. Long-span Contextual Prosodic Cues Extraction

To take the advantage of higher level information and alleviate the interferences of supra-segment- or word-level effects, the differences between the target syllable nucleus and its 2 preceding and 2 successive (in total 4) neighboring syllable nuclei are computed as shown in Figure 5. In this stage, both the pitch and energy feature vectors are 36-dimension and duration vector is 4 dimensions. Final, there are in total 76-dimension prosodic feature vectors.

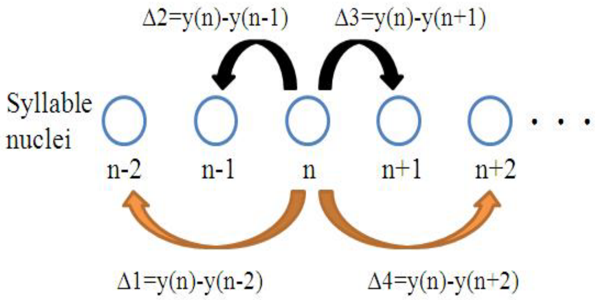


Figure 5: *Schematic diagram of the contextual features extraction.*

3. Prosodic Feature Selection

Although, there are many potential prosodic features listed in Table 1, it is not sure which cues are the most useful ones for stressed/unstressed syllable classification. In order to find the best features, a binary decision tree-based feature selection algorithm is adopted to identify the most informative features. Figure 6 show a typical tree built by the algorithm for energy-related features (see Section 4.3.3) in our experiments. In this study, two feature selection procedures will be applied to pitch and energy, separately, and only the top three features chosen by each decision tree will be picked up (in total 7 dimensions, 3 pitch + 3 energy + 1 duration).

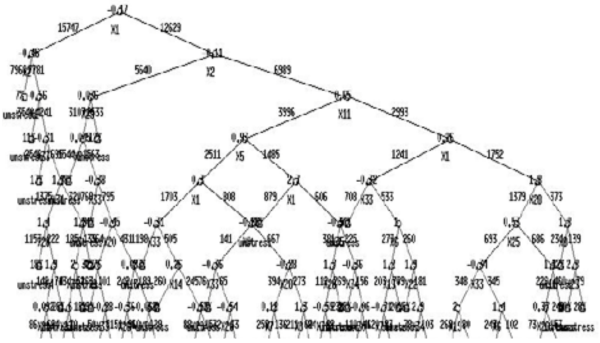


Figure 6: *A typical example of decision tree for energy feature selection (part).*

4. Experiment and Result

The performance of the method was evaluated on "The Kiel Corpus of Read Speech, Vol. I" database. In the following subsections, the corpus and experimental setting are first described and then the performances of the conventional and proposed feature normalization methods are compared using Gaussian mixture model (GMM)- and multi-layer perceptron (MLP)-based classifiers.

4.1. Kiel Corpus

The number of speakers in this corpus is 25 (12 females and 13 males, all are native speakers). There are in total 3,890 sentences with 17,280 words and 37,209 syllables. In average, there are 9.56 syllables for each sentence. Moreover, the ratio between stressed and unstressed syllables is about 4:6. This database was further divided into three disjoint subsets, including a training, a development and an evaluation ones by the ratio of 8:1:1 (28,376 training, 4,392 development and 4,441 test syllables) in our experiments. Moreover, manual phone segmentations are available in this database.

4.2. Experimental Setting

In all the following experiments, Praat [13] was used to extract pitch contour of an input utterance. Fujisaki model Parameter Extraction Environment [14] was adopted to decompose the pitch contour and remove the undesired (Fb) and (Ap) components. On the other hand, energy contour was calculated using root mean square (RMS) formulation and a band-pass filter (bandwidth 0.5~4.5 Hz) was applied. Finally, 19-dimenssion prosodic features were computed and expanded into 76-dimensional long-span contextual prosodic cues using given manual segmentations. Besides, decision tree-based feature selection, GMM and MLP modules from LNKnet Pattern Classification Software [15] were adopted. The parameters of the decision trees, GMMs and MLPs were empirically determined using the development data.

4.3. Experimental Results

Performances of the conventional and proposed method were evaluated and compared using error rate (Err) criterion as defined in Eq. (2). Three scenarios were tested including (1) Z-Score with and without Fujisaki model/band-pass filtering front-end, (2) Z-Score plus Fujisaki model/band-pass filtering front-end methods with and without long-span contextual prosodic cues, and (3) feature selection with and without long-span contextual prosodic cues.

$$Err(\%) = \frac{\#.\text{of classification errors}}{\#.\text{of syllables}} \times 100\% \qquad (2)$$

4.3.1. Z-Score baseline w/o Fujisaki model/filtering

First of all, the whole set of features listed in Table 1, i.e., 19 dimensions, was extracted and normalized using (1) Z-Score only or (2) Fujisaki model/band-pass filtering plus Z-Score methods. The results in Table 2 show that MLPs outperformed GMMs and Fujisaki model/filtering front-end did improve the Err of the MLP recognizer from 31.34% to 27.04%.

4.3.2. Proposed method w/o long-span contextual cues

Secondary, the set of 19 features normalized using Fujisaki model/band-pass filtering plus Z-Score method was expanded into 76-dimensional long-span contextual prosodic cues. The results in Table 3 show that long-span contextual prosodic cues are very informative and further lower the Err of the MLP recognizer from 27.04% to 25.60%.

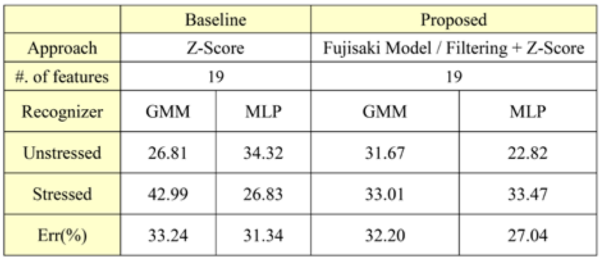
4.3.3. Feature selection w/o long-span contextual cues

Thirdly, from the results of decision tree analysis, it is found that the ranking of pitch-related features is "Span > 1st >

Max". For energy-related ones, the order is "Max > Span > Q3". Therefore, a reduced set of features (7-dimension) was selected and further expanded into 28-dimensional long-span contextual feature vectors.

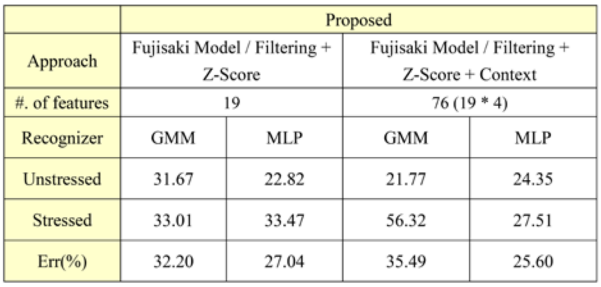
Table 4 shows the results of the proposed feature selection method with and without contextual prosodic cues. It could be found by comparing Table 4 with 3 that the performance of the selected 7-dimensional features is slightly better than the whole feature set. Moreover, from Table 4, the selected contextual features worked much better and achieved the lowest Err of 24.68% using the MLP classifier.

Table 2. *Performance comparison between conventional Z-Score only baseline and the proposed Z-Score plus Fujisaki model/filtering front-end methods.*



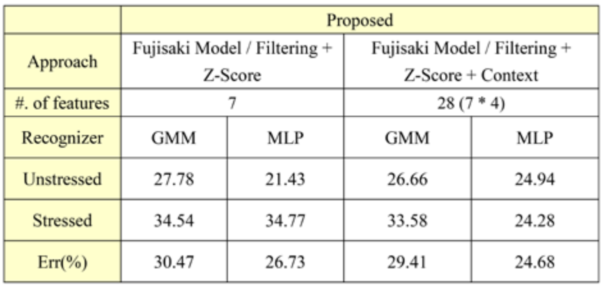
	Baseline		Proposed	
Approach	Z-Score		Fujisaki Model / Filtering + Z-Score	
#. of features	19		19	
Recognizer	GMM	MLP	GMM	MLP
Unstressed	26.81	34.32	31.67	22.82
Stressed	42.99	26.83	33.01	33.47
Err(%)	33.24	31.34	32.20	27.04

Table 3. *Performance of the proposed Z-Score plus Fujisaki model/filtering front-end methods with and without long-span contextual prosodic cues.*



	Proposed			
Approach	Fujisaki Model / Filtering + Z-Score		Fujisaki Model / Filtering + Z-Score + Context	
#. of features	19		76 (19 * 4)	
Recognizer	GMM	MLP	GMM	MLP
Unstressed	31.67	22.82	21.77	24.35
Stressed	33.01	33.47	56.32	27.51
Err(%)	32.20	27.04	35.49	25.60

Table 4. *Performance of the proposed feature selection method with and without long-span contextual prosodic cues.*



	Proposed			
Approach	Fujisaki Model / Filtering + Z-Score		Fujisaki Model / Filtering + Z-Score + Context	
#. of features	7		28 (7 * 4)	
Recognizer	GMM	MLP	GMM	MLP
Unstressed	27.78	21.43	26.66	24.94
Stressed	34.54	34.77	33.58	24.28
Err(%)	30.47	26.73	29.41	24.68

5. Conclusions

A multi-stage feature normalization scheme which takes both word stress and sentence intonation patterns into consideration is proposed in this paper. The performance evaluation results on "The Kiel Corpus of Read Speech, Vol. I" database has shown that Z-Score plus Fujisaki model/filtering front-end worked better than the Z-Score only baseline. Besides, further integration of decision tree-based feature selection and long-span contextual prosodic cues have achieved the lowest Err of 24.68%. These results confirm the efficiency of the proposed approach for German stressed/unstressed syllable classification.

6. Acknowledgements

This work was supported by the National Science Council, Taiwan, under the project with contract NSC 98-2221-E-027-081-MY3 and NSC 100-2221-E-027-006.

7. References

[1] F. Tamburini and C. Caini, "An automatic system for detecting prosodic prominence in American English continuous speech," International Journal of Speech Technology, vol. 8, no. 1, 2005, pp. 33–44.

[2] D. Wang and S. Narayanan, "An acoustic measure for word prominence in spontaneous speech," IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 2, 2007, pp. 690–701.

[3] Q. Shi, et al., "Spoken English Assessment System for Non-Native Speakers Using Acoustic and Prosodic Features," Interspeech, 2010.

[4] H. Fujisaki and K. Hirose, "Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese," Journal of the Acoustical Society of Japan (E) 5 (4), 1984, pp. 233-242.

[5] H. Xie, P. Andreae, M. Zhang, and P. Warren, "Detecting stress in spoken English using decision trees and support vector machines," Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation. 2004, pp. 145–150, Australian Computer Society, Inc.

[6] J. Tepeerman and S. Narayanan, "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Philadelphia, vol. I, 2005, pp. 937-940.

[7] M. Lai, Y. Chen, M. Chu, Y. Zhao and F. Hu, "A Hierarchical Approach to Automatic Stress Detection in English Sentences," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, and ICASSP, vol. I,, 2006, pp. 753-756.

[8] G. M. Chang, "Discriminative Feature Analysis based on Voice Onset Time and Stress Detection for Taiwanese-accented English Speech," Master Thesis of National Cheng Kung University, 2007.

[9] C. Y. Tseng, "An Initial Study on Stress Detection for Spoken English," Master Thesis of National Tsing Hua University, 2008.

[10] G. Senthil Raja and S. Dandapat, "Speaker recognition under stressed condition," Proceedings of International Journal of Speech Technology, 2010, pp. 141–161.

[11] H. Mixdorff, "An Integrated Approach to Modeling German Prosody," Proceedings of Habilitation thesis submitted to TU Dresden, 2002. Vol. 25

[12] The Kiel Corpus, http://www.ipds.uni-kiel.de/forschung/kielcorpus.en.html, 2011

[13] P. Boersma and D. Weenink, "Praat: doing phonetics by computer Toolbox," available from the link at the author's homepage at http://www.fon.hum.uva.nl/praat/, 2011

[14] H. Mixdorff, "Fujisaki model Parameter Extraction Toolbox," available from the link at the author's homepage at http://public.beuth-hochschule.de/~mixdorff/thesis/fujisaki.html, 2011

[15] LNKnet Pattern Classification Software http://www.ll.mit.edu/mission/communications/ist/lnknet/index.html, 2011