THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA VOL.23, NO.4E 2004. 12 pp. 112-119

Automatic Recognition of Pitch Accents Using Time-Delay Recurrent Neural Network

Sung-Suk Kim*, Chul Kim*, Wan-Joo Lee*
*School of Computer & Information, Yong-In University
(Received April 27 2004; accepted December 15 2004)

Abstract

This paper presents a method for the automatic recognition of pitch accents with no prior knowledge about the phonetic content of the signal (no knowledge of word or phoneme boundaries or of phoneme labels). The recognition algorithm used in this paper is a time-delay recurrent neural network (TDRNN). A TDRNN is a neural network classier with two different representations of dynamic context: delayed input nodes allow the representation of an explicit trajectory F0(t), while recurrent nodes provide long-term context information that can be used to normalize the input F0 trajectory. Performance of the TDRNN is compared to the performance of a MLP (multi-layer perceptron) and an HMM (Hidden Markov Model) on the same task. The TDRNN shows the correct recognition of 91.9% of pitch events and 91.0% of pitch non-events, for an average accuracy of 91.5% over both pitch events and non-events. The MLP with contextual input exhibits 85.8%, 85.5%, and 85.6% recognition accuracy respectively, while the HMM shows the correct recognition of 36.8% of pitch events and 87.3% of pitch non-events, for an average accuracy of 62.2% over both pitch events and non-events. These results suggest that the TDRNN architecture is useful for the automatic recognition of pitch accents.

Keywords: *Pitch accent, Prosody recognition, Speech recognition, TDRNN, MLP, HMM*

I. Introduction

Words that a talker considers semantically or pragmatically important are often produced with a fundamental frequency contour called a pitch accent. A pitch accent is an unusually high F0 (possibly a local maximum) or an unusually low F0 (possibly a local minimum) designed to draw attention to the important word[1]. The presence of a pitch accent correlates with other changes in the acoustic signal, including increased duration of vowels and increased burst amplitude and voice onset time (VOT) of stop consonants. Cole et al.[2] found evidence that accentual strengthening of stop consonants shifts the boundary between voiced and unvoiced cognates of any given stop release: for example, the VOT and burst amplitude of an unaccented /p/ release are comparable to the VOT and burst amplitude of an accented /b/ release. Knowledge of pitch accent placement would therefore be useful prior information for automatic speech recognition.

This paper proposes a method for the automatic recognition of pitch accents with no prior knowledge about the phonetic content of the signal (no knowledge of word or phoneme boundaries or of phoneme labels). In the framework presented here, the problem of pitch accent recognition is considered to be a special case of the general problem of context-dependent, non-parametric dynamic contour recognition. The recognition problem is non-parametric because the distribution of F0 is unknown; in particular,

Corresponding author: Sung-Suk Kim (sskim@yongin.ac.kr)
School of Computer & Information, Yong-In University, 470 Samga-dong, Yongin-si, Kyonggi-do, Korea

there is no evidence that the distribution of F0 is Gaussian. The recognition problem is context-dependent because F0 encodes much more than just prosody: in particular, talker dependence, dependence on speaking style, and short-time acoustic phonetic information encoded in the F0 trajectory must be ignored. The recognition algorithm used in this paper is a time-delay recurrent neural network (TDRNN) [3]. A TDRNN is a neural network classier with two different representations of dynamic context: delayed input nodes allow the representation of an explicit trajectory F0(t), while recursive nodes provide long-term context information that can be used to normalize the input F0 trajectory. Section II of this paper describes a selection of papers in the field of prosody dependent speech recognition, and briefly discusses the importance of the problem. Section III describes the TDRNN architecture for pitch accent recognition used in these experiments. Section IV describes the experimental methods used for the recognition of pitch accents. Section V gives the experimental results, and Section VI presents conclusions.

II. Background

Prosodic labels are potentially useful in automatic speech understanding systems for at least four reasons. First, prosody correlates with syntax: Price et al.[4] showed that prosody may be used to disambiguate syntactically distinct sentences with identical phoneme strings, while Kim et al.[5] have demonstrated that prosody may be used to infer punctuation of a recognized text. Second, prosody correlates with meaning: for example, Taylor et al.[6] have used prosody for the purpose of recognizing the dialog act labels of utterances. Third, prosody is useful for the detection and subsequent processing of speech disfluencies[7]. Finally, prosody may be useful as prior conditioning information for the correct phoneme labeling of an ambiguous acoustic signal. The acoustic implementation of a phoneme depends on its prosodic context in many ways: accented vowels tend to be longer and less subject to coarticulatory variation[8], while accented consonants are produced with greater closure duration[9], greater linguopalatal contact[10], longer voice onset time, and greater burst amplitude[2].

In an automatic speech recognition system, prosody may be recognized before, after, or simultaneous with the recognition of phonemes and words. The ordering of the word-recognition module and the prosody-recognition module depends on the intended purpose of prosody recognition. Systems that intend to use prosody only for the purpose of semantic, syntactic, or disfluency processing often implement a prosodic post-processing strategy, in which the input to the prosody recognizer includes a time-aligned word graph generated by an initial prosody-independent speech recognizer. The advantage of a post-processor strategy is greater accuracy, won by the use of syllable-timed acoustic features (e.g., average F0 during the syllable of interest[11]) and word string information [12]. These advantages are compelling in many applications:all reported uses of prosody in commercial speech under- standing systems use a post-processor model of prosody recognition. The disadvantage of a post-processor strategy is that the front-end recognizer is unable to use prosody to aid in the phonetic labeling of ambiguous acoustic signals. Kompe[11] demonstrates both theoretically and empirically that a prosody post-processor can improve the search time of a speech recognizer, but never its word recognition accuracy.

Taylor[13] has demonstrated one of the few systems able to recognize pitch accents without prior information about word boundary location. His two-stage prosody recognition system first locates pitch events using a hidden Markov model, then labels the pitch events using an analysis-by-synthesis matching strategy. "Pitch events" include high, major pitch accents and rising phrase boundaries. Non-events include minor accents, level accents, and falling boundaries, as well as regions with no perceptible prosodic contour. The best reported pitch event recognition system comprises three-state mixture-Gaussian hidden Markov models of each distinct pitch event label, meaning that every accent type, every boundary type, and every possible combination of an accent and a boundary are distinctly modeled. The HMM observes talker-normalized F0 and delta-F0. For the purposes of scoring recognizer output, all pitch event models (accent and rising boundary models) output the label "e" (event), and all non-event models (including level accent, minor accent, falling boundary, and continuation segment) output the label "c." Event recognition is considered correct only if the overlap between a transcribed pitch event and a true pitch event is

at least 50% of the duration of the true pitch event. Under these constraints, speaker-independent pitch event recognition correctness is 72.7%, with a recognition accuracy of 47.7% (25% insertion rate). Speaker-dependent pitch event recognition correctness is 82.1%, with an accuracy of 63.1%.

The symbols of intonational phonology are the subject of current debate, and a variety of annotation systems have been used to transcribe publicly distributed databases. The Boston Radio News corpus[14] is transcribed using the Tones and Break Indices (ToBI) annotation standard [1,15,16]. ToBI posits three fundamental pitch movements: high (H), low (L), and downstepped (!H). A pitch accent is composed of one or two pitch movements in a row; an intermediate phrase boundary tone is a single pitch movement, and an intonational phrase boundary tone is a sequence of two movements. About 95% of pitch accents in the Radio News corpus are centered on a high pitch movement (H* and L+H* accents) or a downstepped pitch movement (!H*, L+!H*, or H+!H*). Dainora[17] argues that !H and H movements are not linguistically distinct and should therefore not be distinctly recognized. The Radio News corpus documentation further notes that the distinction between L+H* and H* is the least reliably transcribed.

Taylor and his colleagues have annotated the DCIEM maptask[18] using an annotation scheme with three types of accent (high, level, and minor, marked as h, l, and m), two types of boundary (rising and falling, marked as rb and fb), and a flat connecting contour (c). Despite the detailed notation, Taylor argues that low pitch events represent a default setting of the speech production mechanism rather than a consciously produced pitch accent; his recognition results are reported for the task of distinguishing pitch "events" (a and rb) from nonevents (l, m, fb, and c). In ToBI notation, Taylor's pitch events correspond approximately to high pitch accents (H*) and rising boundary tones (H-, L-H%, and H-H%) while low accents (L*) and falling boundary tones (L-, H-L%, and L-L%) roughly correspond to pitch non-events.

III. TDRNN Architecture

In this section, we describe a neural network architecture called time-delay recurrent neural network (TDRNN) for automatic recognition of pitch movements. The architecture of the TDRNN is shown in Fig. 1[3]. The TDRNN is a 3 layer back-propagation network with an additional pitch context layer. The TDRNN provides two different representations of dynamic context. First, time-delayed input units (as in a TDNN[19]) allow the representation of short-term context of an explicit trajectory F0(t). Second, multiple recurrent circuits through time-delayed pitch context layer units encode long-term context information that can be used to normalize the input F0 trajectory. The activation of the pitch layer unit at time t is copied into that of the pitch context layer unit which is used for long-term context modelling of pitch movements and acts as an additional input at time $t+1$.

The Delay Box of n interconnections shown in Fig. 1, each with its own time-delay, from the input unit to the hidden unit and between the pitch context unit and the hidden unit can be depicted as Fig. 2. Node i of layer $h-1$ is connected to node j of the next upper layer h, with the connection line having an independent time-delay $\tau_{ijk,h-1}$ and weight $w_{jik,h-1}$. Each node sums up the net inputs from the activation values of the previous layer

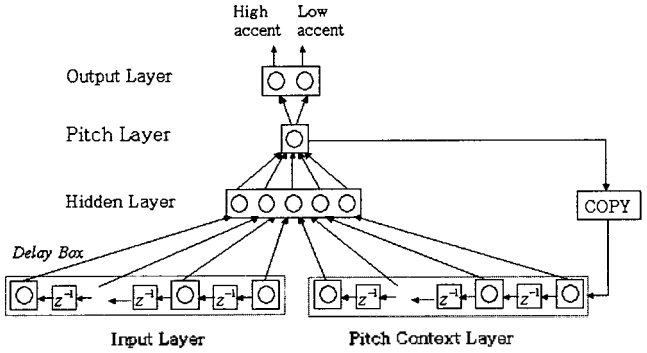

Figure 1. The architecture of TDRNN (z^{-1} denotes 1 time frame delay).

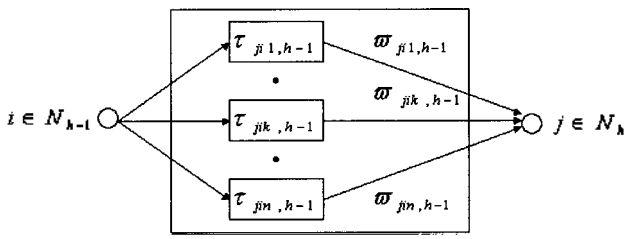

Figure 2. Delay box ($\tau_{ijk,h-1}$ denotes kth time frame delay to node j from node i).

nodes, through the corresponding time-delay on each connection line, i.e., at time t_n, node j of layer h receives a weighted sum:

$$x_{j,h}(t_n) = f\Big(\sum_{i \in N_{h-1}} \sum_{k=1}^{K_{ji,h-1}} w_{jik,h-1} \cdot x_{i,h-1}(t_n - \tau_{ijk,h-1})\Big) \quad (1)$$

where $x_{i,h-1}(t_n - \tau_{ijk,h-1})$ is the activation level of layer of node i on the layer $h-1$ at time $t_n - \tau_{ijk,h-1}$, N_{h-1} denotes the set of nodes of layer $h-1$, $K_{ji,h-1}$ represents the number of connections (i.e., number of time-delays) to node j of layer h from node i of layer $h-1$, and $f(.)$ is a non-decreasing sigmoidal function. The interconnection weights $w_{jik,h-1}$ are learned using the error back-propagation learning algorithm[20], and the time-delays $\tau_{ijk,h-1}$ are fixed. Since the activation of the pitch context layer unit is a direct copy of the previous pitch layer activation, the feedback connection through the copy operation is not subject to training.

IV. Experimental Methods

The TDRNN was trained and tested for the purpose of talker-independent, gender-dependent pitch event recognition using data extracted from the Boston Radio News Corpus [14]. Performance of the TDRNN was compared to the performance of a TDNN/MLP (time-delay neural network/ multi-layer perceptron) and an HMM-based recognizer trained and tested on the same task.

The Boston Radio News Corpus is a series of radio stories read by seven professional radio announcers, and partially annotated using the ToBI (tones and break indices) prosodic annotation system[15, 16]. Seven types of pitch accents are transcribed in the Radio News Corpus. All seven types of accents involve classification of pitch movement on the accented syllable into one of three categories: high (H*), downstepped (!H*), and low(L*). The notation "*?" is used to mark a questionable pitch accent. Some transcriptions mark the location of a pitch accent (as "*") but not its type; most of these are high or downstepped accents. The TDRNN, MLP, and HMM recognizers in our experiments are trained to recognize as pitch events all syllables marked with H*, !H*, *, ?*, or L*, and as non-events all unaccented syllables. For training purposes, each pitch event or non-event starts at the beginning of the first sonorant phoneme in a syllable, and ends at the end of the last sonorant phoneme in the same syllable. For testing purposes, all three recognizers were used to label every frame in the test database as either accented or unaccented. During recognition tests, a pitch event was considered correctly recognized if at least 50% of its frames were labeled "accented." The TDRNN, MLP, and HMM are trained using 67 paragraphs comprising 2,078 pitch events and 2,116 non-events from one female speaker (F1A), and tested with 164 paragraphs comprising 6,999 pitch events and 7,082 non-events from another female speaker (F2B).

Both TDRNN and MLP networks observe a two-dimensional input vector containing normalized fundamental frequency ($\widehat{F_0}(t)$) and energy ($\widehat{E_0}(t)$). The fundamental frequency $F_0(t)$ is extracted using the formant program in Entropic XWAVES with probability of voicing (PV) output as a confidence measure for the extracted $F_0(t)$. We eliminated pitch doubling and halving errors by eliminating F_0 that falls into the doubling and halving clusters of a 3 mixture Gaussian model whose mixture component means are restricted to $1/2\mu$, μ, and 2μ, where μ is the estimated utterance mean F_0. We then normalize F_0 by μ and convert it to log scale:

$$\widehat{F_0}(t) = \max(0.2,\ \log(\frac{F_0(t)}{\mu} + 1)). \quad (2)$$

To eliminate unreliable $\widehat{F_0}$ measures, those with PVs smaller than a heuristic threshold are replaced by the linear interpolated values $\widehat{F_0}$ based on the $\widehat{F_0}$ that have PVs greater than the threshold. Similarly, energy is normalized using:

$$\widehat{E_0}(t) = \max(-3,\ \log(\frac{E_0(t)}{\eta})), \quad (3)$$

where η is the utterance maximum energy.

The TDRNN is configured with 2 input units, 10 hidden

units, 1 pitch layer unit (1 pitch context layer unit), and 2 output units. The input units have 14 time frame delays for input context modeling, while the recursive pitch context layer unit has 18 time frame delays for the long-term context modeling of pitch movements. The MLP is configured with 2 input units, 20 hidden units, and 2 output units. The input units have 16 time frame delays to provide context to the network, and 17 frames are used as input. The HMM-based recognizer uses five three-state HMMs, modeling the five labels H*, !H*, ?*, L*, and unaccented (there were no * labels in the training data). Of several tested HMM configurations, best performance was achieved using a ten-component diagonal-covariance mixture Gaussian PDF with a six-dimensional feature vector comprising $\overline{F}_0(t)$, $\overline{E}_0(t)$, and their deltas and delta-deltas. The HMMs have 393 trainable parameters each, for a total of 2358 trainable parameters. The MLP has 742 trainable parameters (720 weights + 22 biases), and the TDRNN has 505 trainable parameters (492 weights + 13 biases); thus the neural network architectures use 20% to 30% as many parameters as the HMM.

The TDRNN is trained, using error back-propagation, to imitate a target function. The target function for the TDRNN is equal to 1 during pitch events, and 0 otherwise (thus the TDRNN target function is equal to 0 during pitch non-events, and also during non-sonorant frames). The two output units of the MLP are trained, using error back-propagation, to imitate complementary target functions: one is equal to the target function of the TDRNN, while the other is equal to one minus the TDRNN target. The HMMs are trained using a standard Baum-Welch maximum likelihood training algorithm.

In recognition tests, both the TDRNN and the MLP were used to label every frame in the test database as either pitch event or non-event. A pitch event was considered correctly recognized if the recognized pitch event and the true pitch event overlapped in time by at least 50%, as proposed by Taylor[13].

V. Experimental Results

We have performed the testing which is speaker independent, by running the trained TDRNN, MLP, and 5 mixture HMM over test data. The TDRNN has two output units; one represents pitch accent event marked with high accents (H*, !H*, *), and the other represents pitch event marked with low accents (L*, ?*), but the TDRNN creates important information about pitch non-event which can be derived from the pitch layer unit. Fig. 3 shows three functions of time, computed for a sentence in the test database: the TDRNN target function, the TDRNN output unit, and the TDRNN pitch layer unit. As shown, the TDRNN output unit tracks the target function reasonably well. The figure also demonstrates the way in which the pitch layer unit encodes information about the long-term context of the input pitch contours. Specifically, the activation of the pitch layer unit approximates, reasonably well, the complement of the target function: the probability of a pitch non-event.

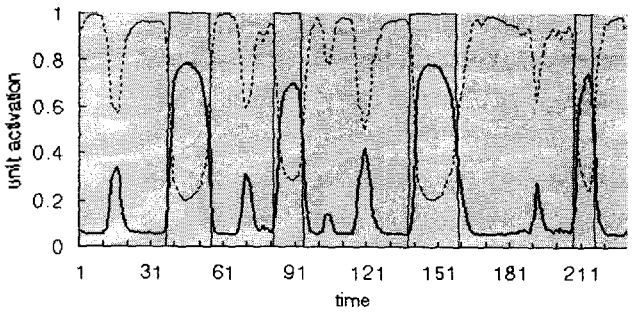

Figure 3. The solid line is for true pitch event, the bold solid line is for the output values of the output layer unit of TDRNN, and the dashed line is for the output values of the pitch layer unit which estimates pitch non-events.

Recognition experiments were performed using test data extracted from the Radio News Corpus. The MLP marked each frame as a pitch event if and only if the activation of the pitch-event output node was higher than that of the non-event node. The TDRNN marked each frame as a pitch event if and only if the activation of the output unit was higher than that of the pitch layer unit. Pitch event recognition accuracy is computed as the number of correctly recognized events divided by the number of events in the database; in order to be correct, the recognized time period of an event must overlap with the true time period of the same event by at least 50%. Non-event recognition accuracy is scored the same way.

Table 1 summarizes pitch event and non-event recognition results on the test data. The TDRNN shows the correct recognition of 91.9% of pitch events and 91.0% of pitch non-events, for an average accuracy of 91.5% over both

Table 1. The results of pitch event recognition with TDRNN, MLP, and HMM.

Architecture	Pitch accent events (%)	Pitch non-events (%)	Pitch accent events + Pitch non-events (%)
TDRNN	91.9	91.0	91.5
MLP	85.8	85.5	85.6
HMM	36.8	87.3	62.2

Table 2. The individual results of pitch event recognition by the TDRNN: the columns show how many tokens were recognized as event, and how many were recognized as non-event.

Pitch accent	Recognized as event Percentage(%)(Number)	Recognized as non-event Percentage(%)(Number)
H*	93.0 (4331)	7.0 (325)
!H*	91.5 (1392)	8.5 (130)
*	94.7 (126)	5.3 (7)
?*	88.7 (337)	11.3 (43)
L*	78.9 (243)	21.1 (65)
Unaccented	9.0 (634)	91.0 (6448)

Table 3. The individual results of pitch event recognition by the MLP: the columns show how many tokens were recognized as event, and how many were recognized as non-event.

Pitch accent	Recognized as event Percentage(%)(Number)	Recognized as non-event Percentage(%)(Number)
H*	87.8 (4088)	12.2 (568)
!H*	87.3 (1329)	12.7 (193)
*	94.0 (125)	6.0 (8)
?*	75.8 (288)	24.2 (92)
L*	57.5 (177)	42.5 (131)
Unaccented	14.5 (1030)	85.5 (6052)

Table 4. The individual results of pitch event recognition by the 5 mixture HMM: the columns show how many tokens were recognized as event, and how many were recognized as non-event.

Pitch accent	Recognized as event Percentage(%)(Number)	Recognized as non-event Percentage(%)(Number)
H*	37.2 (1733)	62.8 (2923)
!H*	36.9 (562)	63.1 (960)
*	44.4 (59)	55.6 (74)
?*	36.8 (140)	63.2 (240)
L*	26.9 (83)	73.1 (225)
Unaccented	12.7 (897)	87.3 (6185)

pitch events and non-events. The MLP with contextual input exhibits 85.8%, 85.56%, and 85.6% recognition accuracy respectively, while the 5 mixture HMM shows the correct recognition of 36.8% of pitch events and 87.3% of pitch non-events, for an average accuracy of 62.2% over both pitch events and non-events. Table 2, Table 3, and Table 4 show how many tokens were recognized as "event," and how many were recognized as "non-event". These results show that the TDRNN encodes dynamic variations of pitch movements better than the others (MLP and HMM) do. The 5 mixture HMM, in particular, almost doesn't recognize pitch accents, while well recognizes pitch non-events.

VI. Conclusions

The results reported in this paper may be meaningfully compared to three sets of prior published results. First, human transcribers agree on the location of pitch accents in the Boston University Radio News corpus with an agreement rate of roughly 91%[14]. Second, Taylor[13] has published the only experimental results for the task of pitch event recognition without given word boundary locations. His mixture-Gaussian HMM-based intonation recognition system achieved 82.1% recognition correctness and 63.1% recognition accuracy on a speaker-dependent subset of the DCIEM map-task corpus (72.7% correctness and 47.7% accuracy on a speaker-independent subset). Third, Ostendorf and Ross[21] have trained and tested a speaker-dependent mixture Gaussian stochastic segment model with word and phoneme boundaries specified by an HMM pre-processor. Their model was both trained and tested using talker F2B from the Radio News Corpus, the same talker that were used to test the MLP and TDRNN models in our experiments. Their model recognizes pitch accent location with an accuracy of 89%. If the results are collapsed into our categories of pitch event and non-event, the resulting accuracies are 78.2% correct recognition of events, and 94.9% correct recognition of non-events.

The recognition accuracy of the talker-independent TDRNN architecture in this paper is almost identical to the rate of agreement among human transcribers (91.5% versus 91%). The TDRNN performs significantly better than accent recognizers based on HMM or stochastic segment models containing mixture Gaussian probability densities; the TDRNN also performs significantly better than a non-recursive MLP architecture.

The relatively good performance of the TDRNN on this

task may be attributed to discriminatively trained non-parametric representation of the long term context-dependent categories of pitch event and pitch non-event. First, the TDRNN represents the discrimination between pitch event and non-event classes using a discriminative non-parametric neural network classifier, instead of using a mixture Gaussian model trained using the maximum likelihood method. Performance of a parametric classifier depends on the details of the probability model in the critical region near the classification threshold; the probability distribution of F0(t) in this critical region may not be well approximated by a mixture Gaussian model. The 5 mixture HMM, in this paper, shows only 62.2% correct recognition accuracy. Second, the TDRNN represents both long-term and short-term context; without long-term context, the MLP architecture achieves 85.6% correct recognition accuracy. These results suggest that the long-term context modeling through multiple recurrent circuits is useful for the correct recognition of pitch events. We will seek to apply the TDRNN for automatic labeling of pitch accents to a prosody dependent speech recognizer that models word and prosody in a unified probabilistic framework.

References

1. Mary E. Beckman and Janet Pierrehumbert. Intonational structure in japanese and english. Phonology Yearbook, 3:255-309, 1986.
2. Jennifer Cole, Hansook Choi, Heejin Kim, and Mark Hasegawa-Johnson. The effect of accent on the acoustic cues to stop voicing in radio news speech. In Proc. Internat. Conf. Phonetic Sciences, 2003.
3. Sung-Suk Kim, Time-delay recurrent neural network for temporal correlations and prediction, Neurocomputing, 20, pp, 253-263, 1998.
4. P.J. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong. The use of prosody in syntactic disambiguation. J. Acoust. Soc. Am, 90 (6) :2956-2970, Dec. 1991.
5. Ji-Hwan Kim and Philip C. Woodland. The use of prosody in a combined system for punctuation generation and speech recognition. In Proc. EUROSPEECH, 2001.
6. P. Taylor, S. King, S. Isard, H. Wright and J. Kowtko, Using intonation to constrain language models in speech recognition, in Proc. EUROSPEECH, 1997.
7. Christine H. Nakatani and Julia Hirschberg. A corpus-based study of repair cues in spontaneous speech. J. Acoust. Soc. Am, 95(3) :1603-1616, 1994.
8. T. Cho. Effects of Prosody on Articulation in English. PhD thesis, UCLA, 2001.
9. Kenneth DeJong. The supraglottal articulation of prominence in english: Linguistic stress as localized hyperarticulation. J. Acoust. Soc. Am, 89(1) :369-382, 1995.
10. *Cecile Fougeron and Patricia Keating.* Articulatory strengthening at edges of prosodic domains. J. Acoust. Soc. Am, 101(6) :3728-3740, 1997.
11. R. Kompe. Prosody in Speech Understanding Systems. Springer-Verlag, 1997.
12. Colin Wightman and Mari Ostendorf. Automatic labeling of prosodic patterns. IEEE Trans. Speech and Audio Processing, 2(4) :469-481, Oct. 1994.
13. Paul Taylor. Analysis and synthesis of intonation using the Tilt model. J. Acoust. Soc. Am, 107(3) :1697-1714, 2000.
14. M. Ostendorf, P.J. Price, and S. Shattuck-Hufnagel. The Boston University Radio News Corpus. Linguistic Data Consortium, 1995.
15. Mary E. Beckman and Gayle M. Ayers. Guidelines for ToBI Labelling: the Very Experimental HTML Version. www.ling.ohiostate.edu/research/phonetics/E ToBI/singer tobi.html, 1994.
16. Joseph F. Pitrelli, Mary Beckman, and Julia Hirschberg. Evaluation of prosodic transcription labeling reliability in the TOBI framework. In Proc. ICSLP, 1994.
17. Audra Dainora. Eliminating downstep in prosodic labeling of american english. In ISCA Workshop on Prosody in Speech Recognition and Understanding, pages 41-46, 2001.
18. E.G. Bard, C. Sotillo, A.H. Anderson, and M.M. Taylor. The DCIEM map task corpus: Spontaneous dialogues under sleep deprivation and drug treatment. In Proc. ESCA-NATO Tutorial and Workshop on Speech under Stress, pages 25-28, Lisbon, 1995.
19. Alexander Waibel, Toshiyuki Hanazawa, Georey Hinton, Kiyohiro Shikano, and Kevin J. Lang. Phoneme recognition using time-delay neural networks. Trans. Acoust. Speech Sig. Proc., 37:328-339, 1989.
20. Rumelhart D. E., McClelland J. L., and the PDP Research Group. Learning representations by back-propagating errors. In Parallel Distributed Processing, 1, pages 318-362. MIT Press, 1986.
21. M. Ostendorf and K. Ross. A multi-level model for recognition of intonation labels. In Computing prosody: computational models for processing spontaneous speech. Springer-Verlag New York, Inc., 1997.

[Profile]

•Sung-Suk Kim

Sung-Suk Kim received the B.S. degree in electrical engineering from Yeungnam University, Gyeongsan, Korea, in 1985, and the M.S. and Ph.D. degrees in electronics and computer engineering from the University of Ulsan, Ulsan, Korea, in 1987 and 1990, respectively. From 1985 to 1991, he was with Korea Electric Power Corporation (KEPCO). He is currently an Associate Professor of School of Computer & Information at Yong-In University, Yongin, Korea. He was a Visiting Researcher at Language Technology Institute, Carnegie Mellon University, Pittsburgh, PA, in 2002, and worked as a Visiting Professor at Beckmann Institute for Advanced Science and Technology, University of Illinois, Urbana-Champaign, IL, in 2003. His research interests includes speech recognition, neural network, computer-assisted language learning (CALL), and blind source separation.

•Chul Kim

Chul Kim received the B.S. degree in electronics engineering, and the Ph.D. degree in computer science from Yonsei University, Seoul, Korea, in 1977 and 2000, respectively. From 1981 to 1983, he was a Researcher at the Samsung Telecommunications Research Institute, Korea. From 1984 to 1993, he also was a Program Manager of Information Technology at IBM Korea Software Development Institute. Since 1994, he has

been with the School of Computer & Information, Yong-In University, Yongin, Korea where he is currently an Associate Professor and he was a Director of the Institute of Natural Sciences from 1997 to 1998. From 1999 to 2000, he was a Visiting Professor of the Department of Computer Science at University of Colorado, Colorado Springs, CO. His research interests include protocol engineering, computer network, and mobile and wireless communications.

•Wan-Joo Lee

Wan-Joo Lee the B.S., M.S., and Ph.D. degrees in electronics engineering from Yonsei University, Seoul, Korea, in 1987, 1989, and 1995, respectively. Since 1995, he has been with the School of Computer & Information, Yong-In University, Yongin, Korea where he is currently an Associate Professor. His research interests are in the areas of image processing and pattern recognition.