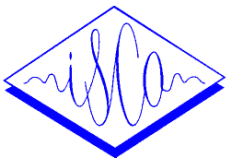


2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

FLEXIBLE VOCABULARY RECOGNITION OF SPEECH

Matthew Lennig, Douglas Sharp, Patrick Kenny, Vishwa Gupta, and Kristin Precoda

Bell-Northern Research and INRS-Télécommunications, 16 Place du Commerce, Montreal, Canada H3E 1H6

ABSTRACT

Speaker-independent speech recognition over the telephone network has begun to be practical for certain simple transactions such as automation of collect and third-number-billed telephone calls (Lennig, 1990). However, the current state of the art requires that such recognizers be trained on the specific vocabulary to be recognized and requires collecting training tokens of each vocabulary item from hundreds or often thousands of different speakers. The goal of the current study is to obviate the need for vocabulary-specific training, thus allowing a new vocabulary to be introduced merely by providing phonemic transcriptions for the words. In this paper we present speaker-independent results over the telephone network in which not only the speakers but also the vocabularies used in the training and test sets are disjunct. The vocabulary consists of the names of the 1,561 companies having common stock listed on the New York Stock Exchange in mid-1986. Without the aid of a language model, we have achieved speaker-independent, vocabulary-independent recognition results of 96% correct.

1. INTRODUCTION

Current-day voice dialog systems that work over the telephone network and employ speaker-independent speech recognition use fixed vocabularies, usually consisting of the digits and some control words. If a particular interactive voice application has a big enough market to warrant the expense, a customized, speaker-independent vocabulary can be constructed by collecting tokens of the words spoken by many people over the telephone network. The number of training tokens required can vary from dozens to thousands, depending on the accuracy requirements of the target system.

For large scale applications such as the automation of collect and third-number-billed telephone calls (Lennig, 1990), the cost of collecting a large number of training tokens can be justified. However, such applications are rare. We contend that the vast majority of potential applications for speaker-independent speech recognition will only become practical when new vocabularies can be created without having to collect training data.

The most demanding of these applications require frequent updating of their vocabularies. The example chosen in this paper is an interactive voice service for dispensing stock quotations. Every week, new companies are listed on the New York Stock Exchange. To add these newly listed company names to the system's vocabulary, the system administrator should simply have to type the names in, along with the phonemic transcriptions and ticker symbols.

The goal of the work described in this paper is *flexible vocabulary recognition,* which we define as the ability to specify a recognizer's vocabulary simply by providing a phonemic transcription for each word. To assess vocabulary flexibility, we have run a series of experiments in which the training and test vocabularies are disjunct. Since these are speaker-independent experiments, the training and test speaker sets are also disjunct.

The first successful use of hidden Markov models (HMM's) to represent phonetic units is reported in Jelinek (1976). INRS-Télécom-munications has extended this work to recognize a vocabulary of 86,000 words (Lennig et al., 1990a; Gupta et al., 1991). As with most large vocabulary phoneme-based recognizers, this result was speaker-dependent, isolated word, and used high quality input speech. The DARPA work (e.g., Kubala et al., 1991) has extended the state of the art of phonemic HMM recognition to continuous speech, speaker-independent recognition, but still employs high quality input speech. No attempt has been made to conduct vocabulary-independent experiments.

Other work in this direction was reported by Hon et al. (1989) and by Hon and Lee (1990, 1991), although they have not worked with telephone speech and no explicit attempt was made to use disjunct training and test set vocabularies.

The work presented in this paper represents a fusion of the speaker-dependent, phoneme-based work at INRS-Telecom and parallel work at Bell-Northern Research on small vocabulary, word-based, speaker-independent recognition (Lennig, 1988 and 1990).

2. RECOGNITION VOCABULARY

The recognition vocabulary consists of the 1,561 names of companies which had common shares listed on the New York Stock Exchange (NYSE) as of mid-1986. The alphabetized list of these 1,561 words[1] was partitioned into training and test vocabularies by assigning alternate words to the training and test set vocabularies, respectively. The training set vocabulary consists of 781 words. The test set vocabulary contains 780 words. No word appears in both training and test vocabularies.

No individual spoke any word more than once. Because the training and test set vocabularies were sampled randomly to generate the training and test material, some words have zero tokens associated with them.

During recognition experiments, all 1,561 words appearing in both training and test sets were in the active vocabulary and could be recognized. No language model was used, and all 1,561 words were assigned equal *a priori* probability. The recognition vocabulary was therefore 1,561 despite the fact that not all 1,561 words have tokens appearing in the test corpus.

3. TRAINING AND TEST SPEECH DATA

A total of 127 English-speaking men and women living in greater Montreal each read a different list of 40 company names over the telephone. They were instructed to read each word naturally. 59 of the 127 speakers were designated as training set speakers and the remaining 68 were designated as test set speakers. Each speaker placed a dialed-up local telephone call to our data collection system and read his or her unique list of 40 company names. The telephone calls were placed from a variety of telephone sets in a variety of environments, typically office environments.

To create the word lists, the 1,561-word vocabulary was first partitioned into two disjoint subsets by alphabetizing the entire list and assigning words alternately to each subset. One subset was arbitrarily designated as the training vocabulary subset and the other as the test vocabulary subset. A word list was created for each training speaker by randomly choosing 40 words (with replacement) from the training vocabulary subset. Similarly, a word list was created for each test speaker by randomly selecting 40 words from the test vocabulary subset. Since the training and test vocabulary

[1]For the purposes of this research, each name was treated as a "word" by the recognizer and will be referred to as such in this paper, even though many company names consist of multiple linguistic words, on average 2.1 linguistic words per company name.

10.21437/ICSLP.1992-26

subsets were disjunct, no word in the vocabulary appeared in both training and test corpora. This is the necessary condition to conduct a vocabulary-independent experiment.

The result is a speech corpus containing 40 random tokens from each of 127 subjects where about half the subjects spoke words from half of the 1,561-word vocabulary and the rest of the subjects spoke words from the remainder of the vocabulary. This is the experimental setup that we need to test speaker-independent, vocabulary-independent recognition of company names over the telephone network.

The training set consisted of 2212 tokens from 59 speakers. For computational reasons, only the first 14 tokens produced by each test set speaker were used in the experiments reported here, resulting in a total of 952 test tokens from 68 speakers. Table 1 summarizes the training and test set statistics.

Table 1. Training and test set statistics. Same set coverage indicates the lexical coverage of the training set vocabulary by the training set tokens and of the test set vocabulary by the test set tokens.

	Speech Corpus	
	Training	**Test**
Vocabulary size	781	780
Number of tokens	2212	952
Same set coverage	98%	66%
Number of speakers	59	68

4. DESCRIPTION OF THE RECOGNIZER

Preprocessing

The mel-frequency cepstral coefficients and their first differences make up the 15-dimensional feature vector computed each frame as described in Davis and Mermelstein (1980) and Lennig et al. (1990b) and summarized below.

Incoming speech is low-pass filtered and sampled using a standard telephone codec chip. The resulting 8 kHz 8-bit μ-law samples are converted to linear samples and blocked into 204-sample analysis windows. The window is advanced by 102 samples every frame to produce an output parameter vector every 12.75 ms. A 204-point Hamming window is applied to the speech samples and a 256-point FFT with zero-sample padding is performed on the frame to produce a 128-point power spectrum. In each of 20 mel-frequency bands, spaced linearly from zero to 1,000 Hz and exponentially from 1,000 Hz to 4,000 Hz, the power spectrum points are combined using a weighted average to simulate the output of a triangular filter. The 20 outputs next undergo a logarithm transformation followed by a cosine transform to yield the mel-frequency cepstrum, (C_1,...,C_7). C_0 is computed as a perceptually weighted sum of the 20 log outputs to deemphasize the contribution of the first few outputs. First differences for (C_0,...,C_7) are computed from the frames 25.5 ms ahead and 25.5 ms behind the current frame by taking the signed vector difference of these two frames. Combining static and dynamic parameters, each frame gives rise to 15 parameters of the form (C_1,...,C_7,ΔC_0,...,ΔC_7).

Phoneme models

The recognizer uses phonemic hidden Markov models, one model for each of the phonemes used to represent North American English. The HMM's are left-to-right models in which three transitions are possible from each state: self-loop, go-to-next-state, and skip-next-state. Associated with each transition is an output distribution on the event space of parameter frames and a transition probability. These parameters are trained using the Viterbi algorithm.

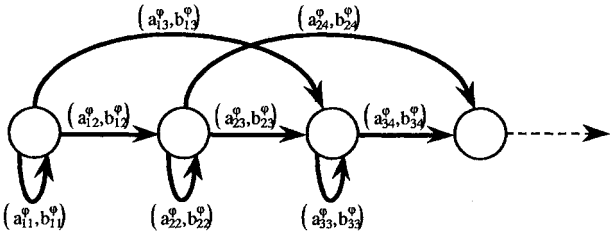
The HMM topology of a five state model is shown in Figure 1, where the dashed line represents a null transition having a transition probability of unity and no output frames.

The mixture densities are computed as follows:

$$b_{ij}^{\varphi}(\mathbf{x}) = \sum_{k=1}^{12} \frac{w_{ijk}^{\varphi}}{(2\pi)^{d/2}|C_{\varphi}|^{1/2}} e^{-1/2\,(\mathbf{x}-\mu_{ijk}^{\varphi})^{t}\, C_{\varphi}^{-1}(\mathbf{x}-\mu_{ijk}^{\varphi})}$$

where w_{ijk}^{φ} and μ_{ijk}^{φ} are the weight and mean of the kth mixture component on the transition from state i to state j, $\mathbf{x}$ is the observed parameter vector, φ is the phoneme index, and C_{φ} the covariance matrix associated with phoneme φ.

Figure 1. HMM structure used to model a phoneme. The number of states may vary (see Table 2) depending on which phoneme is being modeled.



One hidden Markov model is trained for each of the following 39 phonemes:

/ptkbdgθfsʃðvzʒʧʤrljwmnŋhieEaˈaʷɔˈɑoʊɪɛæʊʌə/

In addition, models for a short schwa allophone [ᵊ], interword silence, pre-utterance silence, and post-utterance breath/silence are trained, making a total of 43 HMM's. The number of states in each model, and an example word illustrating each phoneme, is given in Table 2.

Output distributions

The output distributions are modeled as continuous output mixture gaussians with a full covariance matrix. One covariance matrix per HMM is used, representing the pooled covariance matrix for all output distributions in the model. Twelve gaussian mixture components are used. Thus, for a model having N transitions, one NxN symmetric matrix and $12N$ 15-dimensional vectors are trained.

Training and recognition algorithms

The training set was used to estimate parameters for the phoneme models using a training method employing energy and duration constraints as described in Gupta et al. (1991). Recognition was performed using the Viterbi algorithm.

5. RESULTS

Two forced-choice experiments were run. In the first experiment, only the mel-based cepstrum coefficients and their first differences were used. This gave an error rate of 5%. A complete list of (substitution) errors for this experiment is shown in Table 3.

In a second experiment, two recognizers were run on each utterance: one using mel-frequency cepstral coefficients and the other using line spectrum pairs (Soong and Juang, 1984). The recognizer using line spectrum pairs (LSP's) used the first seven LSP's, their first dynamic parameter counterparts, and dynamic C_0, giving the following 15-dimensional feature set:

$(\omega_1,...\omega_7,\Delta\omega_1,...\Delta\omega_7,\Delta C0)$, where the ω_i represent the LSP parameters. The top choice from each recognizer was then scored by the other recognizer and the log likelihoods added. This reduced the error rate to 4%.

6. REAL TIME IMPLEMENTATION

The recognizer was implemented in near real time on a Sky Challenger board in a Motorola VME system. In order to meet the speed requirements, we used the triphone heuristic A* search described in Kenny et al. (1991, 1993).

Table 2. Phoneme models: example words and number of states

phoneme	example word	number of states
/p/	**p**ie	5
/t/	**t**ea	5
/k/	**k**ey	6
/b/	**b**oat	4
/d/	**d**i**d**	5
/g/	**g**a**g**	4
/θ/	**th**in	5
/f/	**f**ast	5
/s/	**s**auce	5
/ʃ/	**sh**y	5
/ð/	**th**ey	5
/v/	**v**ery	5
/z/	**z**inc	5
/ʒ/	vi**s**ion	5
/tʃ/	**ch**ap	5
/dʒ/	**j**oke	3
/ɹ/	**r**ay	5
/l/	**l**ow	5
/j/	**y**es	5
/w/	**w**ay	5
/m/	**m**et	5
/n/	**n**o**n**e	5
/ŋ/	si**ng**	5
/h/	**h**ello	5
/i/	b**ee**	6
/e/	s**ay**	10
/E/	v**a**ry, v**e**ry	6
/aʲ/	b**uy**	10
/aʷ/	c**ow**	5
/ɔʲ/	t**oy**	5
/ɑ/	f**a**ther	6
/o/	g**o**	6
/u/	d**o**	10
/ɪ/	p**i**ck	6
/ɛ/	p**e**ck	6
/æ/	p**a**ck	6
/ʊ/	b**oo**k	5
/ʌ/	p**u**ck	6
/ə/	**a**bout	5
[°]	butt**o**n	5
preutterance silence		5
interword silence		5
postutterance silence		5

7. STOCK QUOTE APPLICATION

The flexible vocabulary recognition (FVR) technology described above has been implemented in StockTalk, an interactive voice application which dispenses real time stock quotes by voice over the telephone. StockTalk can be accessed from any telephone by dialing 1-800-661-STOCK. It has been in operation since May 1992 as a research trial. A block diagram of StockTalk is shown in Figure 2.

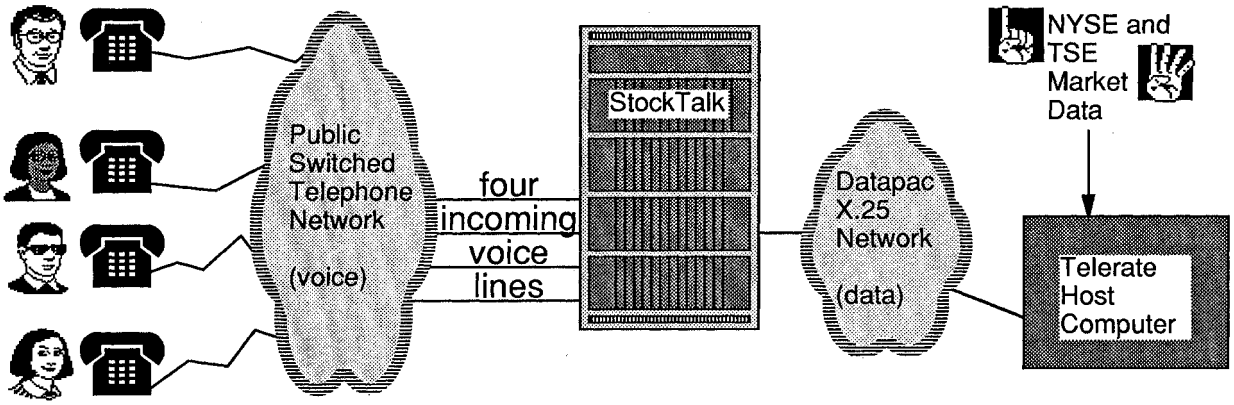
Callers may request a quote on a company by saying its name. The system recognizes all 1,991 company names (as of 13 July 92) having common stock listed on the New York Stock Exchange. We update this vocabulary weekly simply by providing phonemic transcriptions of newly listed company names along with their ticker symbols. In July 1992, StockTalk also began providing real time quotes for the 883 common stocks listed on the Toronto Stock Exchange (TSE).

When the user selects the New York Stock Exchange, the entire (currently) 1,991-word vocabulary becomes active, plus a dozen control words. All 2,003 words are treated by the recognizer as

Table 3. Complete list of errors for the first experiment, in which cepstrum and Δ-cepstrum parameters were used to recognize a 952-token test set with a 1,561-word vocabulary.

Word Pronounced	Word Recognized
Fleming	Grumman
Lehman	3M
Southeastern Public Service	Pacific Resources
Radice	Reece
Radice	Korea Fund
Knight Ridder *(pron. Knight Rider)*	Stride Rite
Hitachi	Apache
ALLTEL	Contel
Squibb	Jack Winter
Savin	Salomon
Sabine	Teledyne
Barnes Group	Anacomp
Timeplex	Banc One
Allied Supermarkets	Allied Signal
Clayton Homes	KeyCorp
Goodyear	Christiana
Katy Industries	JP Industries
Sears Roebuck	Unocal
LeaRonal	Fluor
Questar	Bell & Howell
MAXXAM Group	Marcade Group
Seagram	CIGNA
Airlease	Celanese
Pennzoil	Van Dorn
Inco	Dayco
Zemex	Pandick
Reese	Grainger
Squibb	Twin Disk
Daniel Industries	Bell Industries
Mercury Savings and Loan	American Savings & Loan
Motorola	Bausch & Lomb
Brown Group	Ranco
Lehman	Mead
Gleason	Korea Fund
Southdown	Pulte Home
Roper	Whirlpool
Brown Group	Cannon Group
Oakwood Homes	Alcoa
Thrifty	Certain-teed
Wyle Laboratories	Warner Lambert
Zenith Labs	Ingersoll Rand
Sony	Fleming
Winners	Harris
Sony	Tonka
Raytheon	Christiana
Federated Dept Stores	Century Telephone Enterprises
Thackeray	Barclay's
Unitrode	Eaton
GenCorp	General Homes

Figure 2. Configuration of StockTalk, a real time stock quote system which uses flexible vocabulary recognition to recognize 2,000 names of New York Stock Exchange issues.



having equal *a priori* probability. No attempt is made at this point to perform recognition rejection: recognition is done on a forced-choice basis.

Despite somewhat higher error rates in the field, users are extremely happy with StockTalk. Cooperative users are nearly always able to obtain the quotes they desire.

8. SUMMARY AND CONCLUSIONS

Vocabulary-independent, speaker-independent experiments have been performed over the public switched telephone network on a 1,561 vocabulary of company names with uniform prior probabilities resulting in a 5% error rate for cepstrum (and their first differences) alone and 4% when combined with LSP's. In both cases, training and test set vocabularies were constructed to be disjunct in order to prove the feasibility of over-the-telephone, speaker-independent, flexible vocabulary recognition. A small number of acoustic units (39 phonemes and one allophone) was used to model the vocabulary.

Flexible vocabulary recognition has been put to practical use in StockTalk, a real time stock quote system. New issues are listed on the New York Stock Exchange every week, necessitating additions to the StockTalk vocabulary. We are able to add these company names to StockTalk's vocabulary simply by specifying their phonemic transcriptions. No spoken tokens of the new words are required. StockTalk would be impractical to maintain using conventional speaker-independent speech recognition technology, which requires hundreds or frequently thousands of tokens of each new vocabulary item to be spoken by a variety of speakers.

Finally, the low error rates using only 39 phonemic units point to the usefulness of the phoneme as a fundamental unit for speech recognition.

REFERENCES

Davis, S.B. and P. Mermelstein (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing,* **ASSP-28**(4), 357-366.

Gupta, V.N., M. Lennig, P. Mermelstein, P. Kenny, F. Seitz, and D. O'Shaughnessy (1991). Using phoneme duration and energy contour information to improve large vocabulary isolated word recognition. *Proceedings of the 1991 IEEE International Conference on Acoustics, Speech, and Signal Processing,* 341-344.

Hon, H., K. Lee, and R. Weide (1989). Towards speech recognition without vocabulary-specific training. *Proceedings of Eurospeech.*

Hon, H. and K. Lee (1990). On vocabulary independent speech modeling. *Proceedings of the 1990 IEEE International Conference on Acoustics, Speech, and Signal Processing,* 725-728.

Hon, H. and K. Lee (1991). Recent progress in robust vocabulary-independent speech recognition. *Proceedings of the Fourth DARPA workshop on speech and Natural Language,* Asilomar, 19-22 February 1991.

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE* **64**, 532-556.

Kenny, P., R. Hollan, V. Gupta, M. Lennig, P. Mermelstein, and D. O'Shaughnessy (1991). A*-admissible heuristics for rapid lexical access. *Proceedings of the 1991 IEEE International Conference on Acoustics, Speech, and Signal Processing,* 689-692.

Kenny, P., R. Hollan, V. Gupta, M. Lennig, P. Mermelstein, and D. O'Shaughnessy (1993). "A*-admissible heuristics for rapid lexical access," to appear in *IEEE Transactions on Speech and Audio Processing,* Vol. 1, No. 1, January 1993.

Kubala, F., S. Austin, C. Barry, J. Makhoul, P. Placeway, R. Schwartz (1991). BYBLOS speech recognition benchmark results. *Proceedings of the Fourth DARPA Workshop on Speech and Natural Language Processing.*

Lennig, M. and P. Mermelstein (1988). First public trial of a speech-recognition-based 976 directory. *Proceedings of Speech Tech '88,* New York City, 26-28 April 1988, 291-292.

Lennig, M. (1990). Putting speech recognition to work in the telephone network. *Computer,* August 1990, 35-41.

Lennig, M., V. Gupta, P. Kenny, P. Mermelstein, and D. O'Shaughnessy (1990a). An 86,000-Word Recognizer Based on Phonemic Models. *Proceedings of the DARPA Speech and Natural Language Workshop,* Los Altos: Morgan Kaufmann, 391-396.

Lennig, M., P. Mermelstein, and V.N. Gupta (1990b). Speech Recognition. United States Patent No. 4,956,865, issued September 11, 1990.

Soong, F. and B. Juang (1984). Line spectrum pair (LSP) and speech data compression. *Proceedings of the 1984 IEEE International Conference on Acoustics, Speech, and Signal Processing,* 1.10.1-1.10.4.