HIGH PERFORMANCE "GENERAL PURPOSE" PHONETIC RECOGNITION FOR ITALIAN

Piero Cosi and John-Paul Hosom***

*Istituto di Fonetica e Dialettologia -- C. N. R., Via G. Anghinoni, 10 - 35121 Padova ITALY
email: cosi@csrf.pd.cnr.it

**Center for Spoken Language Understanding (CSLU-OGI)
Oregon Graduate Institute (OGI), P.O. Box 91000, Portland, Oregon 97291 USA
email: hosom@cse.ogi.edu

ABSTRACT

The development of a speaker independent "general purpose" phonetic recognizer for Italian is described. The CSLU Toolkit was used to develop and implement the system. The recognizer, based on a frame-based hybrid HMM/ANN architecture trained on context-dependent categories to account for coarticulatory variation, recognizes 38 different phonemes (not including silence or closures), and can distinguish between stressed and unstressed vowels as well as open and closed vowels. The APASCI corpus, containing nearly 2500 sentences read by 100 speakers, where the sentences have been designed to maximize the number of phonemes occurring in different contexts, was used for training and testing. As of the time of this writing, a phoneme-level accuracy of 82.90% on the development set and of 80.53% on the test set has been obtained. This level of accuracy is much greater than on a similar English-language corpus (with state-of-the-art performance of slightly better than 70%) and it represents the best performance obtained so far on this corpus.

1. INTRODUCTION

In many tasks a speaker-independent domain-specific vocabulary (such as "collect call", "calling card", "operator", or "help") needs to be recognized. For such tasks, a general-purpose (***gp***) recognizer that is capable of recognizing all permissible phoneme strings in a language is required. Moreover, such a recognizer would be of great help in supporting research and development of new spoken dialogue systems and to exploit new dialogue strategies by the use of specific software such as the Rapid Application Developer included in the CSLU Speech Toolkit [1]

2. CSLU SPEECH TOOLKIT

The recognizer being described in this work was developed and implemented by the use of the CSLU Speech Toolkit [1]. Since the CSLU Toolkit has been described in several recent articles [1, 2, 3, 4] and is available through the CSLU OGI Web site [5], we limit our discussion to only a brief overview of the main toolkit components. The CSLU Speech Toolkit is a comprehensive set of tools and technologies for learning about, researching and developing interactive language systems and their underlying technologies. The Toolkit supports real-time interactive dialogues on standard off-the-shelf PC platforms running Windows (a Linux version will be available soon). It provides a modular, open architecture supporting distributed, cross-platform, client/server-based networking. This flexible environment makes it possible to easily integrate new components and to develop scalable, portable speech-related applications. The components of the Toolkit include both neural-network and HMM-based speech recognition systems, a natural-language semantic parser called PROFER [6], the Festival text-to-speech system [7], an anatomically accurate talking face called Baldi [8], and software for recording, displaying, labeling, and manipulating speech. The Toolkit also includes a GUI-based application developer called RAD and the documentation required to train HMM and neural-network based recognizers. The tools are designed to enable inexperienced users to rapidly design, test and deploy spoken language systems. In addition to the pre-existing components, users can write their own C-level or script code for tools. Because the Toolkit is portable, runs on affordable off-the-shelf computing platforms, and provides both the knowledge (tutorials) and resources needed to conduct research, it removes some of the main entry barriers that currently prevent universities and research laboratories from establishing new programs in human language technology.

3. RECOGNITION FRAMEWORK

The basic framework for the CSLU Speech Toolkit's hybrid HMM/ANN speech recognition systems is illustrated in Figures 1 and 2. These systems use features that represent the spectral envelope (warped to emphasize the perceptually relevant aspects [9]) and its energy given a fixed window size. These spectral features are computed at every 10-msec frame in the utterance and are input to the neural network for classification. The neural network receives not just the features for a given frame, but a set of features for the given frame and a fixed, small number of surrounding frames. This "context window" of features is used to provide the network with information about the dynamics of the speech signal. At each frame, the neural network classifies the features in the context window into phonetic-based categories, estimating the probabilities of each category being represented by that set of features. The result of the neural network processing is a CxF matrix of probabilities, where C is the number of phonetic-based categories, and F is the number of frames in the utterance. The word or words that best match this matrix of probabilities is determined using a Viterbi search, given the vocabulary and grammar constraints. The search is usually thought of as traversing a state sequence (illustrated in Figure 2 with a simple two-word vocabulary),

where each state represents a phonetic based category, and there are certain probabilities of transitioning from one state to another.

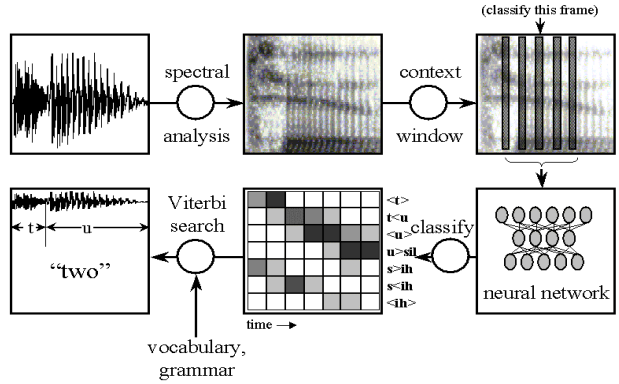

Figure 1. Graphical overview of the recognition process, illustrating recognition of the word "two".

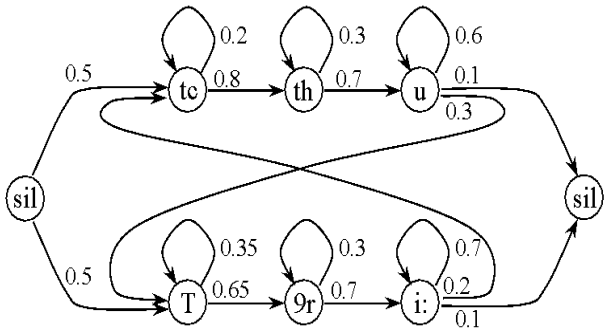

Figure 2. HMM state sequence for a two-word vocabulary.

The major difference between this framework and standard HMM systems is that the phonetic likelihoods are estimated using a neural network instead of a mixture of gaussians. Using a neural network to do this estimation has the advantage of not requiring assumptions about the distribution or independence of the input data, and neural networks easily perform discriminative training [10]. Also, neural networks can be used to perform recognition much faster than standard HMMs. A second difference is in the type context-dependent units. Whereas standard HMMs train on the context of the preceding and following phonemes, our system splits each phoneme into states that are dependent on the left or right context, or are context independent.

4. CORPUS

The Acoustic-Phonetic and Spontaneous Speech Corpus of IRST (APASCI) [11] distributed by ELRA [12] was used for training, development, and testing the general purpose Italian phonetic recognizer. The APASCI corpus was designed and collected at the *Istituto per la Ricerca Scientifica e Tecnologica (ITC/IRST - Trento, Italy)*. This acoustic-phonetic corpus contains Italian read utterances acquired from 100 speakers, 50 females and 50 males. Most of the speakers are from the North-East of Italy. Recordings were performed in a quiet room. Speech was acquired at 48 kHz, with 16 bit accuracy, by means of a digital audio Tape-Recorder Sony TCD-D10PRO and a super-cardioid microphone Sennheiser MKH 416-T. Then digital recordings were downsampled to 16 kHz and speech waveform files (files with extension ".wav") were stored in the SPHERE format [13]. To reduce frequency components under 50 Hz, each signal in the corpus was also high-pass filtered. An artificial grammar was used to randomly generate several sentences, starting from a vocabulary formed by the 1000 most frequent Italian words and other words covering a large variety of phonetic contexts. The choice of grammar ensured that the sentences were syntactically correct, even if often meaningless. A sub-optimal procedure was then used to select a subset of sentences having a good phonetic and "diphonic" coverage. In this way phonetic coverage, in terms of a given number of occurrences, was ensured for the Italian phonemes and pairs of phonemes. Time-aligned phonetic and word transcriptions are provided for each utterance in the corpus. Phonetic transcriptions are given in terms of the Speech Assessment Methods Phonetic Alphabet (SAMPA) [14]. Time-aligned phonetic and word transcriptions were automatically produced by the use of a context-independent continuous density Hidden Markov Model (HMM) automatic speech recognizer and were checked manually.

5. EXPERIMENT

5.1 Acoustic units

The recognizer uses a frame-based hybrid HMM/ANN architecture trained on context-dependent categories to account for coarticulatory variation, recognizes 38 different Italian phonemes (not including silence or closures), and can distinguish between stressed and unstressed vowels as well as open and closed vowels. A three-layer neural network was trained to estimate, at every 10-msec frame, the probability of 545 context-dependent phonetic categories. These categories are created by splitting each phoneme, as illustrated in Table 1, into one, two, or three parts, depending on the length of the phoneme and how much the phoneme was thought to be influenced by coarticulatory effects.

Acoustic units	parts	description
.pau	1	silence
i e E a O o u	3	unstressed vowels
ii ee EE aa OO oo uu	3	stressed vowels
pcl bcl tcl dcl kcl gcl	1	closure
p b t d k g	r	plosive
ts dz dZ tS	2	affricate
s z f v S	2	fricative
m n N	2	nasal
l r L	2	liquid
j w	2	glide
@sch	2	schwa

Table 1. Acoustic units (SAMPA, except closures) and number of parts to split each unit into, for the Italian "general purpose" recognizer (r means "*right dependent unit*").

Phoneme states were trained for different preceding and following phonetic contexts, and some phonetic contexts were grouped together to form a broad-context grouping. The broad-context groupings were done based on acoustic-phonetic knowledge.

Group	Acoustic units in group	Description
$sil	.pau .garbage	silence
$fnt	i ii e ee j	front
$mid	E EE a aa @sch	mid
$bck	O OO o oo u uu w	back
$lab	p b f v m pcl bcl	labial
$alv	t d ts dz s z n tcl dcl	alveolar
$pal	dZ tS S N L	palatal
$vel	k g kcl gcl	velar
$lat	l	lateral
$ret	r	retroflex

Table 2. Groupings of acoustic units into clusters.

5.2 Feature extraction

As for feature extraction, 13 MFCC [9] features (12 cepstral coefficients and 1 energy parameter) plus their delta values are continuously computed with a 10-msec frame rate. Cepstral-mean subtraction (CMS) [15] was performed, with the mean computed using all frames of data. The input to the network consisted of the features for the frame to be classified, as well as the features for frames at -60, -30, 30, and 60 msec relative to the frame to be classified (for a total of 130 input values).

5.3 Training strategy

Neural-network training was done with standard back-propagation on a fully connected feed-forward network. The training was adjusted to use the negative penalty modification proposed by Wei and van Vuuren [16]. With this method, the non-uniform distribution of context-dependent classes, that is dependent on the order of words in the training database, is compensated for by flattening the class priors of infrequently occurring classes. This compensation allows better modeling for an utterance in which the order of the words can not be predicted.

5.4 Duration constraints

Transition probabilities were set to be all equally likely, so that no assumptions were made about the a priori likelihood of one category following another category. In order to make use of a priori information about phonetic durations, and to minimize the insertion of very short words, the search was constrained by specifying minimum duration values for each category, where the minimum value for a category was computed as the value at the second percentile of all duration values. During the search, hypothesized category durations less than the minimum value were penalized by a value proportional to the difference between the minimum duration and the proposed duration. The grammar allowed any phoneme in any order, with optional silence between phonemes.

5.5 "Baseline"

The system was trained, developed and tested with the APASCI corpus. In particular, 1250 hand-labeled sentences were used for training, 105 were considered for the development stage and 715 for the test phase. The training data were searched to find all the vectors of each category in the hand-labeled training section of APASCI. The neural network was trained using the back-propagation method with 130 inputs, 250 nodes in the single hidden layer, and one node for each context-dependent category in the output layer (for a total of 545 output nodes). Training was done for 45 iterations, and the "best" network iteration ("*baseline*" network - ***B***) was determined by phone-level evaluation on the APASCI development-set data.

5.6 "Forced alignment"

Each waveform in the same hand-labeled APASCI training set was then recognized using this B network, with the result constrained to be the correct utterance. This process, called "*forced alignment*", was used to generate time-aligned category labels. These force-aligned category labels were then used in a second cycle of training, which was done again for 45 iterations, and evaluation was repeated to determine the final network ("*force aligned*" network - ***FA***), which was finally evaluated with the APASCI testing speech material.

5.7 "Forward Backward" training

In order to explore the possibility to further improve the recognition results, the "*forward-backward*" (***FB***) training strategy could be [17] applied. Like most of the other hybrid systems, the neural network in this system is used as a state emission probability estimator. A three-layer fully connected neural network can be conceived, with the same configuration as that of the baseline and forced-aligned neural networks and the same output categories. Unlike most of the existing hybrid systems which do not explicitly train the within-phone relative likelihoods, this new hybrid trains the within-phone models to probability estimates obtained from the forward-backward algorithm, rather than binary targets. To start FB training an initial binary-target neural network is required. For this initial network, the best network resulting from forced-alignment training (FA) should be used. Then the forward-backward re-estimation algorithm could be used to regenerate the targets for the training utterances. The re-estimation can be implemented in an embedded form, which concatenates the phone models in the input utterance into a "big" model and re-estimates the parameters based on the whole input utterance. The networks would be trained using the standard stochastic back-propagation algorithm, with mean-square-error as the cost function.

5.8 Results

As of the time of this writing, two of the three stages have been completed: *baseline* and *force alignment* training. As illustrated in Table 3, phoneme-level accuracy of 82.90 and 80.53% on the APASCI development and test set respectively has been obtained.

	Itr #	Snts #	Wrds #	Sub %	Ins %	Del %	PhnAcc %
dev	**24**	**105**	**5235**	**10.41**	**2.56**	**4.45**	**82.90**
test	**24**	**715**	**36439**	**11.97**	**3.24**	**5.12**	**80.53**

Table 3. Recognition performance in terms of phone accuracy for the development and test set.

This level of accuracy is much greater than on a similar English-language corpus (with state-of-the-art performance of slightly

better than 70%) and it represents the best performance obtained so far on this corpus, with no grammar and no phonotactic constraints. In fact, the performance obtained so far by IRST on an extended version of the same APASCI corpus [18] range, at the phone level, from 71.34% to 79.04%, while considering context-independent units (CIUs) and from 75.38% to 76.60% with Syllable-type units (SUs). When context-dependent units (CDUs) were considered, results were slightly better than ours, ranging from 81.36% to 82.44%. However in this case, in contrast with our present implementation, phonotactic constraints were introduced in order to inhibit the recognition of unit sequences having incompatible contexts and this, according to the authors, improved accuracy from 2% to 3% depending on the particular unit set. Moreover in this case a very complex and sophisticated HMM system, with 16 gaussian mixtures per state and a large number (from 337 to 849) of context-dependent states was used in comparison to the rather straightforward architecture of the system being described in this work.

6. CONCLUSIONS

High performance recognition accuracy was achieved with two out of three training stages (*B, FA*) on the APASCI corpus. The third stage of training is currently under development, and test-set evaluation of the system will be subsequently performed. The current-best recognizer was implemented in the Toolkit's dialogue design module and a simple Italian-language demonstration program that accepts menu orders from a user has been created. This demonstration system was installed on a laptop machine and was successful in informal presentations. The present Italian "general purpose" recognizer will be included in the next version of the CSLU Speech Toolkit.

7. ACKNOWLEDGMENT

The authors would like to thank various people from IRST, and in particular Daniele Falavigna, for their always-useful suggestions and important comments. This work was sponsored in part by Consiglio Nazionale delle Ricerche under "International Short-Term Mobility Program".

8. REFERENCES

[1] Sutton S., Novick D.G., Cole R.A. and Fanty M., "Building 10,000 spoken-dialogue systems". In *Proceedings ICSLP96*, Philadelphia, PA, October 1996, Vol. 2, pp. 709-712.

[2] Sutton S., Cole R.A., de Villiers J., Schalkwyk J., Vermeulen P., Macon M., Yan Y., Kaiser E., Rundle B., Shobaki K., Hosom J.P., Kain A., Wouters J., Massaro D., and Cohen M., "Universal Speech Tools: the CSLU Toolkit". In *Proceedings ICSLP98*, Sydney, Australia, November 1998, Vol. 7, pp. 3221-3224.

[3] Cole R.A., Sutton S., Yan Y., Vermeulen P. and Fanty M., "Accessible technology for interactive systems: A new approach to spoken language research". In *Proceedings ICASSP98*, Seattle, Washington, May 1998, Vol. 2, pp. 1037-1040.

[4] Cole R.A., "Tools for research and education in speech science." In *Proceedings ICPhS99*, San Francisco, CA, Aug 1999, Vol. ", pp. 1277-1280.

[5] http://cslu.cse.ogi.edu/toolkit.

[6] Kaiser E.C., Johnston M., and Heeman P.A., "PROFER: Predictive, Robust Finite-State Parsing for Spoken Language". In *Proceedings ICASSP-99*, Phoenix, AZ, March 1999, Vol. 2, pp. 629-632.

[7] Black A. and Taylor P., "Festival Speech Synthesis System: System Documentation (1.1.1)". *Human Communication Research Centre Technical Report HCRC/TR-83*, Edinburgh, 1997.

[8] Massaro D.W., *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. MIT Press: Cambridge, MA, 1998.

[9] Davis S. and Mermelstein P., "Comparison of Parametric Representations for Monosyllabic Word Recognition". *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1980, vol. ASSP-28, pp. 357-366.

[10] Bourlard H., "Towards Increasing Speech Recognition Error Rates". In *Proceedings EUROSPEECH95*, Madrid, Spain, September 1995, Vol. 2, pp. 883-894.

[11] Angelini B., Brugnara F., Falavigna D., Giuliani D., Gretter R. and Omologo M., "Automatic Segmentation and Labeling of English and Italian Speech Databases". In *Proceedings EUROSPEECH93*, Berlin, Germany, 1993, Vol. 1, pp. 653-656.

[12] http://www.icp.grenet.fr/ELRA/cata/spee_det.html#apasci.

[13] ftp://jaguar.ncsl.nist.gov/pub.

[14] Fourcin A.J., Harland G., Barry W. and Hazan W., Eds. *Speech Input and Output Assessment, Multilingual Methods and Standards*, Ellis Horwood Books in Information Technology, 1989.

[15] Furui S., "Cepstral Analysis Techniques for Automatic Speaker Verification". *IEEE Transactions on Acoustic Speech and Signal Processing*, Vol. 29, No. 2, 254-272.

[16] Wei, W. and Van Vuuren, S., "Improved Neural Network Training of Inter-Word Context Units for Connected Digit Recognition". In *Proceedings ICASSP98*, Seattle, Washington, May 1998, Vol. 1, pp. 497-500.

[17] Yan, Y., Fanty, M. and Cole, R., "Speech Recognition Using Neural Networks with Forward-Backward Probability Generated Targets". In *Proceedings ICASSP97*, April 1997, Vol. 4, pp. 3241-3244.

[18] Angelini B., Brugnara F., Falavigna D., Giuliani D., Gretter R., Omologo M., "Speaker Independent Continuous Speech Recognition Using an Acoustic-Phonetic Italian Corpus". In *Proceedings ICSLP94*, Yokohama, Japan, 1994, Vol. 3, pp. 1391-1394.