14

Phone Recognition on the TIMIT Database

Carla Lopes[1,2] and Fernando Perdigão[1,3]
[1]*Instituto de Telecomunicações,*
[2]*Instituto Politécnico de Leiria,*
[3]*Universidade de Coimbra*
Portugal

1. Introduction

In the information age, computer applications have become part of modern life and this has in turn encouraged the expectations of friendly interaction with them. Speech, as "the" communication mode, has seen the successful development of quite a number of applications using automatic speech recognition (ASR), including command and control, dictation, dialog systems for people with impairments, translation, etc. But the actual challenge goes beyond the use of speech in control applications or to access information. The goal is to use speech as an information source, competing, for example, with text online. Since the technology supporting computer applications is highly dependent on the performance of the ASR system, research into ASR is still an active topic, as is shown by the range of research directions suggested in (Baker et al., 2009a, 2009b).

Automatic speech recognition – the recognition of the information embedded in a speech signal and its transcription in terms of a set of characters, (Junqua & Haton, 1996) – has been object of intensive research for more than four decades, achieving notable results. It is only to be expected that speech recognition advances make spoken language as convenient and accessible as online text when the recognizers reach error rates near zero. But while digit recognition has already reached a rate of 99.6%, (Li, 2008), the same cannot be said of phone recognition, for which the best rates are still under 80% [1],(Mohamed et al., 2011; Siniscalchi et al., 2007).

Speech recognition based on phones is very attractive since it is inherently free from vocabulary limitations. Large Vocabulary ASR (LVASR) systems' performance depends on the quality of the phone recognizer. That is why research teams continue developing phone recognizers, in order to enhance their performance as much as possible. Phone recognition is, in fact, a recurrent problem for the speech recognition community.

Phone recognition can be found in a wide range of applications. In addition to typical LVASR systems like (Morris & Fosler-Lussier, 2008; Scanlon et al., 2007; Schwarz, 2008), it can be found in applications related to keyword detection, (Schwarz, 2008), language recognition, (Matejka, 2009; Schwarz, 2008), speaker identification, (Furui, 2005) and applications for music identification and translation, (Fujihara & Goto, 2008; Gruhne et al., 2007).

The challenge of building robust acoustic models involves applying good training algorithms to a suitable set of data. The database defines the units that can be trained and

[1] Phone recognition using TIMIT Database, [9]

the success of the training algorithms is highly dependent on the quality and detail of the annotation of those units. Many databases are insufficiently annotated and only a few of them include labels at the phone level. So the reason why the TIMIT database (Garofolo et al., 1990) has become the database most widely used by the phone recognition research community is mainly because it is totally and manually annotated at the phone level.

Phone recognition in TIMIT has more than two decades of intense research behind it and its performance has naturally improved with time. There is a full array of systems, but with regard to evaluation they concentrate on three domains: phone segmentation, phone classification and phone recognition. While the first reaches rates of 93% [2], (Hosom, 2009), the second reaches around 83% (Karsmakers et al., 2007) and the third stays at roughly 79%, (Mohamed et al., 2011; Siniscalchi et al., 2007). Phone segmentation is a process of finding the boundaries of a sequence of known phones in a spoken utterance. Determining boundaries at phone level is a difficult problem because of coarticulation effects, where adjacent phones influence each other. Phonetic classification is an artificial but instructive problem in ASR, (Sha & Saul, 2006). It takes the correctly segmented signal, but with unknown labels for the segments. The problem is to correctly identify the phones in those segments. Phone models compete against each other in an attempt to set their label to the respective segment. The label of the winning model is compared with the corresponding TIMIT label and a hit or an error occurs. Nevertheless, phone classification allows a good evaluation of the quality of the acoustic modelling, since it computes the performance of the recognizer without the use of any kind of grammar, (Reynolds & Antoniou, 2003). Phone recognition obeys harder and more complex criteria. The speech given to the recognizer corresponds to the whole utterance. The phone models plus a Viterbi decoding find the best sequence of labels for the input utterance. In this case a grammar can be used. The best sequence of phones found by the Viterbi path is compared with the reference (the TIMIT manual labels for the same utterance) using a dynamic programming algorithm, usually the Levenshtein distance, which takes into account phone hits, substitutions, deletions and insertions.

The use of hidden Markov models (HMMs) is widespread in speech recognizers, at least for event time modelling. After decades of intensive research everything indicates that the performance of HMM-based ASR systems has reached stability. In the late 1980s artificial neural networks (ANNs) (re)appeared as an alternative to HMMs. Hybrid HMM/ANN methods emerged and achieved results comparable, and sometimes superior, to those of HMMs. In the last decade two new techniques have appeared in the machine learning field, with surprising results in classification tasks: support vector machines (SVMs) and, more recently, conditional random fields (CRFs). But the best results in TIMIT are achieved with hybrid ANN/HMM models, (Rose & Momayyez, 2007; Scanlon et al., 2007; Siniscalchi et al., 2007), and hybrid CRF/HMM models, (Morris & Fosler-Lussier, 2008).

This chapter will focus on the TIMIT phone recognition task and cover issues like the technology involved, the features used, the TIMIT phone set, and so on. It starts by describing the database before looking at the state-of-art regarding the relevant research on the TIMIT phone recognition task. The chapter ends with a comparative analysis of the milestones in phone recognition using the TIMIT database and some thoughts on possible future developments.

[2] Boundary agreement within 20 ms

2. TIMIT Acoustic-Phonetic Continuous Speech Corpus

The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT - Texas Instruments (TI) and Massachusetts Institute of Technology (MIT)), (Garofolo et al., 1990), described in (Zue et al., 1990), contains recordings of phonetically-balanced prompted English speech. It was recorded using a Sennheiser close-talking microphone at 16 kHz rate with 16 bit sample resolution. TIMIT contains a total of 6300 sentences (5.4 hours), consisting of 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. All sentences were manually segmented at the phone level.
The prompts for the 6300 utterances consist of 2 dialect sentences (SA), 450 phonetically compact sentences (SX) and 1890 phonetically-diverse sentences (SI).
TIMIT Corpus documentation suggests training (≈ 70%) and test sets, as described in Table *1*. The training set contains 4620 utterances, but usually only SI and SX sentences are used, resulting in 3696 sentences from 462 speakers. The test set contains 1344 utterances from 168 speakers. The core test set, which is the abridged version of the complete testing set, consists of 192 utterances, 8 from each of 24 speakers (2 males and 1 female from each dialect region). With the exception of SA sentences which are usually excluded from tests, the training and test sets do not overlap.

Set	# speakers	#sentences	#hours
Training	462	3696	3.14
Core test	24	192	0.16
Complete test set	168	1344	0.81

Table 1. TIMIT Corpus training and test sets

TIMIT original transcriptions are based on 61 phones, presented in Table 2. The alphabet used – TIMITBET – was inspired by ARPABET. Details of transcription and manual alignment can be found in (Zue & Seneff, 1996) and phonetic analysis in (Keating et al., 1994). The 61 TIMIT phones are sometimes considered a too narrow description for practical use, and for training some authors compact the 61 phones into 48 phones. For evaluation purposes, the 61 TIMIT labels are typically collapsed into a set of 39 phones, as proposed by Lee and Hon, (Lee & Hon, 1989).
This speech corpus has been a standard database for the speech recognition community for several decades and is still widely used today, for both speech and speaker recognition experiments. This is not only because each utterance is phonetically hand labelled and provided with codes for speaker number, gender and dialect region, but also because it is considered small enough to guarantee a relatively fast turnaround time for complete experiments and large enough to demonstrate systems' capabilities.

2.1 Standard evaluation phone recognition metrics

In ASR systems, the most common phone recognition evaluation measures are phone error rate (PER), or the related performance metric, phone accuracy rate. The latter is defined by the following expression:

$$Accuracy = \frac{(N_T - S - D - I)}{N_T} \times 100\% \quad (1)$$

	Phone Label	Example		Phone Label	Example		Phone Label	Example
1	iy	b*ee*t	22	ch	*ch*oke	43	en	butt*on*
2	ih	b*i*t	23	b	*b*ee	44	eng	Wash*ing*ton
3	eh	b*e*t	24	d	*d*ay	45	l	*l*ay
4	ey	b*ai*t	25	g	*g*ay	46	r	*r*ay
5	ae	bat	26	p	*p*ea	47	w	*w*ay
6	aa	b*o*b	27	t	*t*ea	48	y	*y*acht
7	aw	b*ou*t	28	k	*k*ey	49	hh	*h*ay
8	ay	b*i*te	29	dx	mu*dd*y	50	hv	a*h*ead
9	ah	b*u*t	30	s	*s*ea	51	el	bott*le*
10	ao	b*ough*t	31	sh	*sh*e	52	bcl	b closure
11	oy	b*oy*	32	z	*z*one	53	dcl	d closure
12	ow	b*oa*t	33	zh	a*z*ure	54	gcl	g closure
13	uh	b*oo*k	34	f	*f*in	55	pcl	p closure
14	uw	b*oo*t	35	th	*th*in	56	tcl	t closure
15	ux	t*oo*t	36	v	*v*an	57	kcl	k closure
16	er	b*ir*d	37	dh	*th*en	58	q	glotal stop
17	ax	*a*bout	38	m	*m*o*m*	59	pau	pause
18	ix	deb*i*t	39	n	*n*oo*n*	60	epi	epenthetic silence
19	axr	butt*er*	40	ng	si*ng*			
20	ax-h	s*u*spect	41	em	bott*om*	61	h#	begin/end marker
21	jh	*j*oke	42	nx	wi*nn*er			

Table 2. 61 TIMIT original phone set.

where N_T is the total number of labels in the reference utterance and S, D and I are the substitution, deletion and insertion errors, respectively. $PER = 100\% - Accuracy$. Another measure is correctness, which is similar to accuracy, but where insertion errors are not considered. The number of insertion, deletion and substitution errors is computed using the best alignment between two token sequences: the manually aligned (reference) and the recognized (test). An alignment resulting from search strategies based on dynamic programming is normally used successfully for a large number of speech recognition tasks, (Ney & Ortmanns, 2000). Speech recognition toolkits, such as HTK, (Young et al., 2006), include tools to compute accuracy and related measures on the basis of the transcribed data and recognition outputs using this dynamic programming algorithm.

3. Overview of current and past research on TIMIT phone recognition task

In spite of the advances made in recent decades, the TIMIT phone recognition task is still a challenging and difficult task. Many attempts have been made to improve phone recognizer performance, including the use of better features or multiple feature sets, improved statistical models, training criteria, pronunciation modelling, acoustic modelling, noise handling, language modelling, and others.

Regarding the approaches of statistical models, they can be broadly placed in two main categories: generative and discriminative. Phone recognition consists of finding the best possible sequence of phones (Ph) that fit a given input speech X. It is a search problem involving finding the optimal phone sequence Ph^* given by

$$Ph^* = \arg\max_{Ph} \mathrm{P}(Ph \mid X) \tag{2}$$

Generative approaches apply Bayes rule on

arriving to $Ph^* = \arg\max_{Ph} \mathrm{P}(X \mid Ph)\mathrm{P}(Ph)$. This expression relies on a learned model of the conditional probability distribution of the observed acoustic features *X*, given the corresponding phone class membership. The name 'generative' came about because the model "generates" input observations in an attempt to fit the model *Ph*. Generative approaches are those involving HMMs, segmental HMMs, hidden trajectory models, Gaussian mixture models (GMMs), stochastic segment models, Bayesian networks, Markov random fields, etc. The probabilistic generative models based on maximum likelihood have long been the most widely used in ASR. The major advantage of generative learning is that it is relatively easy to exploit inherent dependency or various relationships of data by imposing all kinds of structure constraints on generative learning, (Jiang, 2010).

In contrast, discriminative approaches, such as those based on maximum entropy models, logistic regression, neural networks (multi-layer perceptron (MLP), time-delay neural networks (TDNN) or Boltzmann machines), support vector machines (SVMs) and conditional random fields (CRFs), instead of modelling the distribution of the input data assuming a target class, aim to model the posterior class distributions, maximizing the discrimination between acoustically similar targets.

The relevant research on TIMIT phone recognition over the past years will be addressed by trying to cover this wide range of technologies.

One of the first proposals involving phone recognition on the TIMIT database was presented by Lee and Hon, (Lee & Hon, 1989), just after TIMIT was released in December 1988. Their system is based on discrete-HMMs. The best results were achieved with phones being modelled by means of 1450 diphones (right–context) using a bigram language model. Three codebooks of 256 prototype vectors of linear prediction cepstral coefficients were used as features. They achieved a correctness rate of 73.80% and an accuracy rate of 66.08% using 160 utterances from one test set (TID7). They propose that their results should become a TIMIT phone recognition benchmark. In fact, their paper has become a benchmark not only because of the performance but also because of the phone folding they proposed. These authors folded the 61 TIMIT labels into 48 phones for training purposes. For evaluation purposes, they collapsed the 61 TIMIT labels into 39 phones, which has become the standard for evaluation. Table 3 describes this folding process and the resultant 39 phone set. The phones in the left column are folded into the right column's labels. 23 phone labels disappear and the label "sil" is added to the set. The remaining phones from the original 61-set are left intact.

Also in 1989, Steve Young presented the first version of HTK (hidden Markov model toolkit), (Young et al., 2006). This software package, developed in Cambridge University, allows the construction and manipulation of hidden Markov models and lead to a notable increase in the area of speech recognition. In (Young, 1992) the author presents the concept of HMM state tying using triphone models (left and right context). The goal is to produce a compact set of context dependent HMMs, showing that state tying significantly reduces the number of physical triphone models in training. They generate triphones from a phone set with 48 elements. The experimental conditions are similar to those established by Lee and

Hon (Lee & Hon, 1989), except that they used standard Mel-frequency cepstral coefficients (MFCCs) features and log energy and their first order regression coefficients (deltas - Δ). The best results presented are 73.7% for correctness and 59.9% for accuracy, using the 39 phone set proposed in (Lee & Hon, 1989) and 160 sentences randomly taken from the test set.

aa, ao	aa
ah, ax, ax-h	ah
er, axr	er
hh, hv	hh
ih, ix	ih
l, el	l
m, em	m
n, en, nx	n
ng, eng	ng
sh, zh	sh
uw, ux	uw
pcl, tcl, kcl, bcl, dcl, gcl, h#, pau, epi	sil
q	-

Table 3. Mapping from 61 classes to 39 classes, as proposed by Lee and Hon, (Lee & Hon, 1989). The phones in the left column are folded into the labels of the right column. The remaining phones are left intact. The phone 'q' is discarded.

In 1991 Robinson and Fallside (Robinson & Fallside, 1991) developed a phone recognition system using a recurrent error propagation network that achieved an astonishing result: 76.4% for correctness and 68.9% for accuracy using the same Lee and Hon evaluation set (Lee & Hon, 1989). These results rise to 76.5% and 69.8% for correctness and accuracy, using the complete test set. The authors point out even higher rates (71.2% for accuracy), but the set of phones is no longer the traditional 39; they used a set of 50 phones. In 1993, (Robinson & Fallside, 1991) Robinson et al coupled the recurrent network with an HMM decoder, where the network is used for HMM state posterior probability estimation. This system was tested with the Wall Street Journal database. The TIMIT results came from a hybrid RNN/HMM in 1994, (Robinson, 1994). The inputs to the neural network are features extracted using a long left context. The network is trained using a softmax output under a cross-entropy criterion. The network outputs were trained as a function of the 61 original TIMIT labels. Results regarding the 39 classical phone set achieved 78.6% for correctness and 75% for accuracy. This result is still above recent publications! The paper also presents an interesting comparison of several works on the phone recognition task.

In 1993 Lamel and Gauvain (Lamel & Gauvain, 1993) reported their research on speaker-independent phone recognition using continuous density HMMs (CDHMM) for context-dependent phone models trained with maximum likelihood and maximum a posteriori (MAP) estimation techniques. The feature set includes cepstral coefficients derived from linear prediction coefficients (LPC) plus Δ and ΔΔ cepstrum (second order regression coefficients). Using the complete test set the results were 77.5%/72.9% (correctness /accuracy).

Halberstadt and Glass (Halberstadt & Glass, 1998), as a result of PhD research, (Halberstadt, 1998) proposed a system in 1998 where several classifiers are combined. The training was performed to maximize the acoustic modelling via multiple heterogeneous acoustic measurements. Each classifier is responsible for identifying a subset of the original TIMIT

labels. Separately, 6 classifiers train 60 TIMIT phone labels (they do not consider the glottal stop /q/). There are 3 additional classifiers combining the information from previous classifiers. Table 4. shows the phones trained in each classifier.

Phone Class	# TIMIT labels	TIMIT labels
Vowel/Semivowel (VS)	25	aa ae ah ao aw ax axh axr ay eh er ey ih ix iy ow oy uh uw ux el l r w y
Nasal/Flap (NF)	8	em en eng m n ng nx dx
Strong Fricative (SF)	6	s z sh zh ch jh
Weak Fricative (WF)	6	v f dh th hh hv
Stop (ST)	6	b d g p t k
Closure (CL)	9	bcl dcl gcl pcl tcl kcl epi pau h#
Sonorant (SON)	33	Vowel/Semivowel + Nasal/Flap
Obstruent (OBS)	18	Strong Fric + Weak Fric + Stop
Silence (SIL)	9	Same as Closure

Table 4. Broad classes of phones used in the multiple classifier system proposed in (Halberstadt, 1998).

Classification uses the SUMMIT[3] segment-based recognizer, (Robinson et al., 1993). Gaussian mixture models are used and different phone sets use different features: MFCCs, perceptual linear prediction cepstral coefficients, and a third MFCC-like representation that the authors call "discrete cosine transforms coefficients". They also used windows of different lengths, temporal features, deltas, etc. Phone recognition is achieved by means of two different approaches: one hierarchical and another parallel. The results exceeded those of all the systems existing at the time. Accuracy, using the core test set, reached 75.6%! The long list of tests performed allowed the authors to conclude that better results are achieved using combinations of classifiers trained separately, rather than a single classifier trained to distinguish all the phones or using all the classifiers. The best results achieved were given by a combination of only five of the eight classifiers available.

In 2003 Reynolds and Antoniou (Reynolds & Antoniou, 2003) proposed training a modular MLP. On a first level they trained the 39 phones but used different feature sets (MFCCs, perceptual linear prediction coefficients, LPC and combinations of them). As a result, they collected several predictions for the same phone that are later combined in another MLP. The best results were achieved by optimizing the number of hidden nodes and also using information from seven broad classes, whose composition is shown in Table 5.

In the detection of these broad classes, several context sizes of the input features were tested, and a context of 35 frames (350ms) was found to be the best. With a slightly different test set (they took the core test set out of the complete test set), they report an accuracy of 75.8% for the 39 TIMIT standard phone test set. The paper also gives a good overview of prior work on TIMIT phone recognition and classification.

[3] SUMMIT is a speech recognition system developed at MIT. This speech recognizer uses a landmark-based approach for modelling acoustic-phonetic events and uses finite-state transducer (FST) technology to efficiently represent all aspects of the speech hierarchy including: phonological rules, lexicon, and probabilistic language model.

Phone Class	# TIMIT labels	TIMIT labels
Plosives	8	b d g p t k jh ch
Fricatives	8	s sh z f th v dh h
Nasals	3	m n ng
Semi-vowels	5	l r er w y
Vowels	8	iy ih eh ae aa ah uh uw
Diphthongs	5	ey aw ay oy ow
Closures	2	sil dx

Table 5. Broad classes of phones used in the system proposed by Reynolds and Antoniou, (Reynolds & Antoniou, 2003).

Sha and Saul (Sha & Saul, 2006) present a system which, while its performance is not very competitive, does introduce an interesting idea. They trained GMMs discriminatively, using the SVM's basic principle: attempt to maximize the margin between classes. With MFCCs and deltas as features and using 16 Gaussian mixtures they achieved an accuracy rate of 69.9%.
The result of a study undertaken at Brno University on the use of TRAPs (TempoRAl Patterns) was a paper on the hierarchical structures of neural networks for phone recognition (Schwarz et al., 2006). The focus was to exploit the contribution that the temporal context can make to phone recognition. The system relies on two main lines:

- The TRAP system – a set of MLPs where each neural network receives features of a single critical band as input. The TRAP input feature vector describes the temporal evolution of critical band spectral densities within a single critical band. The MLPs are trained so as to classify the input patterns in terms of phone probabilities. The phone probabilities of all these MLPs (one for each critical band) are given to another MLP – a probability merger – whose output gives a final posterior probability of each phone.
- Temporal context split system– also based on MLPs, assumes that two parts of a phone may be processed independently: one considering left context and the other right context. Two MLPs are trained to produce phone posteriors for left and right contexts. The outputs of these MLPs feed another MLP whose outputs give a final posterior probability for each phone.

The authors compared several input feature sets, networks with outputs giving posterior probabilities of phones and HMMs states, and also tried to find the best number of frequency bands to analyze. The best result achieved 75.16% accuracy. Tuning the number of the MLP's hidden nodes; using a bigram language model and using 5 context blocks (instead of only left and right) they reached an interesting improvement (4.5% relative), resulting in 78.52% accuracy.
The Brno recognizer is based on 39 phones, which are not exactly the standard TIMIT phones. Closures were merged with their burst instead of with silence (bcl b → b) suggesting that it is more appropriate for features which use a longer temporal context such as theirs. Looking at the utterance transcriptions in 87% of the [bcl] occurrences the closure is followed by [b], but not in the other 13% (e.g.: bcl t, bcl el, bcl ix, etc), and the same happens with the other closures. The paper does not make it clear if the closures also merge with the following phone in these situations. Because these speech units are not the standard TIMIT, some authors argue that their results would be probably worse if they use the standard speech units, (Mohamed et al., 2011).

Interesting results are reported by a Microsoft research group devoted to the study of hidden trajectory models (HTM). Deng et al's HTMs are a type of probabilistic generative model which aims to model the speech signal dynamics and add long-contextual-span capabilities that are missing in the hidden Markov models (Deng el al., 2005). A detailed description of the long-contextual-span of hidden trajectory model of speech can be found in (Dong el al., 2006). The model likelihood score for the observed speech data is computed from the estimate of the probabilistic speech data trajectories for a given hypothesized phone sequence, which is given by a bi-directional filter. The highest likelihood phone sequence is found through the A* based lattice search. A rescoring algorithm was specially developed for HTM. In (Deng & Yu, 2007), the results reached 75.17% for accuracy and 78.40% for correctness. Joint static cepstra and their deltas are used as acoustic features by the HTM model.

Rose and Momayyez, (Rose & Momayyez, 2007), use the outputs of eight phonological feature detectors to produce sets of features to feed HMM recognizers. The detectors are time delay neural networks whose inputs are standard MFCC features, with deltas and delta-deltas. The HMM recognizers defined over the phonological feature streams are combined with HMMs defined over standard MFCC acoustic features through a lattice rescoring procedure. For the complete test set they achieved an accuracy of 72.2%.

Knowing that phone confusions occur within similar phones (Halberstadt & Glass, 1998), Scanlon, Ellis and Reilly (Scanlon et al., 2007) propose a system where information coming from a base system is combined with information coming from a set of broad phone class experts (broad phonetic groups). The base system is a hybrid MLP/HMM using PLP features with 1st and 2nd derivatives. The MLP is trained to discriminate the 61 original TIMIT phone set. The broad phonetic groups are presented in Table 6. They trained only four networks' experts: vowels (25 phones), stops (8 phones), fricatives (10 phones) and nasals (7 phones).

Since each broad-class phone's characteristics are quite different, they use, in each MLP expert, different sets of features found by Mutual Information criteria. The number of outputs of each MLP network is the same as the number of TIMIT labelled phones of the corresponding broad phonetic group.

The output of a broad phonetic group detector (also an MLP which, for each frame, gives a probability of the frame belong to a group) is combined with the output of a phone classifier. If they agree (the phone recognized by the phone classifier belongs to the broad phonetic group given by the broad class detector) they patch the phone posteriors given from the broad phonetic group detector onto the phone classifier predictions. This merged information is then given to an HMM decoder. After tuning the system they achieved 74.2% accuracy for the 39 TIMIT standard phone set, using the complete TIMIT test.

Broad Phonetic Groups	TIMIT – labelled phones
Vowels	aa, ae, ah, ao, ax, ax-h, axr, ay, aw, eh, el, er, ey, ih, ix, iy, l, ow, oy, r, uh, uw, ux, w, y
Stops	p, t, k, b, d, g, jh, ch
Fricatives	s, sh, z, zh, f, th, v, dh, hh, hv
Nasals	m, em, n, nx, ng, eng, en
Silences	h#, epi, pau, bcl, dcl, gcl, pcl, tcl, kcl, q, dx

Table 6. Broad classes of phones used in the system proposed by Scanlon, Ellis and Reilly (Scanlon et al., 2007).

In 2004, a 4-institute research project in the ASR field, named ASAT (automatic speech attribute transcription), (Lee et al., 2007) generated several ideas for the phone recognition task, (Morris & Fosler-Lussier, 2006, 2007, 2008; Bromberg et al., 2007). The main goal of ASAT is to promote the development of new approaches based on the detection of speech attributes and knowledge integration. In 2007 in a joint paper (Bromberg et al., 2007) , several approaches are presented on the detection of speech attributes. The overall system contains a front-end whose output gives predictions for the detected attributes as a probability. This front-end is followed by a merger, which combines predictions of several speech attributes and whose output is given to a phone based HMM decoder.

Methods of Detection	Front-end Processing (Features)	Speech Attributes Detected
MLP (Sound Pattern of English)	13 MFCCs 10ms frames	vocalic, consonantal, high, back, low, anterior, coronal, round, tense, voice, continuant, nasal, strident, silence. (14 attributes)
SVM	13 MFCCs 9 context frames 10ms frames	coronal, dental, fricative, glottal, high, labial, low, mid, nasal, round minus, round plus, silence, stop, velar, voiced minus, voiced plus, vowel. (17 attributes)
HMM	13 MFCCs+Δ+ΔΔ 10ms frames	
Multi-class MLPs	13 PLPs+Δ+ΔΔ 9 context frames 10ms frames	Sonority: obstruent, silence, sonorant, syllabic, vowel; Voicing: voiced, voiceless, NA; Manner: approximant, flap, fricative, nasal, flap, stop-closure, stop, NA; Place: alveolar, dental, glottal, labial, lateral, palatal, rhotic, velar, NA; Height: high, low-high, low, mid-high, mid, NA; Backness: back, back-front, central, front, NA; Roundness: nonround, nonround-round, round-nonround, round, NA; Tenseness: lax, tense, NA. (44 atributes)

Table 7. ASAT project (Bromberg et al., 2007): features used in the front-end and the speech attributes detected as a function of the detection method.

The acoustic-phonetic attribute detectors were achieved using several technologies: MLPs, SVMs, HMMs, TDNNs. Depending on the classifier, different sets of features were used (MFCCs, PLPs, and derivatives). The set of attributes also differs in each classifier, in number and in the detected acoustic-phonetic feature. Table 7 shows the features used in the front-end and the detected speech attributes as a function of detection method in the ASAT project.

In order to provide higher-level evidence of use for speech recognition, the attributes were combined. Conditional Random Fields (CRFs) and knowledge-based rescoring of phone lattices were used to combine the framewise detection scores for TIMIT phone recognition.
Several configurations of speech attribute detectors as inputs to the CRF were tested. The best result was achieved combining 44 MLP attribute predictions with 17 HMM predictions - 69.52% accuracy and 73.39% correctness.
Morris and Fosler-Lussier in (Morris & Fosler-Lussier, 2006) used eight MLPs to extract 44 phonetic attributes as depicted in Table 8. After decorrelating these 44 features with a Karhunen-Loeve transform, they are modelled by conventional HMMs with Gaussian mixtures and by CRFs. The best results came from a TANDEM architecture (attributes are used as input features for the HMMs) with triphones modelled with 4 Gaussian mixtures: 72.52%/66.69% (correctness/accuracy). With the CRF system the performance is a bit lower 66.74%/65.23% (correctness/accuracy), but better than the TANDEM HMM with monophones modelled by a single Gaussian.
The same authors published another work in 2008 (Morris & Fosler-Lussier, 2008), where the TANDEM architecture combined with the use of triphones trained with 16 Gaussian mixtures increased accuracy to 68.53%. The best results using the core test set is 70.74% for accuracy, and using a set of 118 speakers (speakers in the core test set as well as the rest of the speakers from the TIMIT test set that are not among the speakers in the development set) the reported accuracy rate rose to 71.49%. These results were attained using CRFs with 105 input features: 61 corresponding to the posterior probabilities of the TIMIT phones given by a single MLP classifier and the remaining 44 features are phonetic attributes originating in 8 MLP classifiers of the phonetic classes described in Table 8. All MLP classifiers were trained using PLPs plus their deltas and delta-deltas.

Attribute	**Possible output values**
sonority	vowel, obstruent, sonorant, syllabic, silence
voice	voiced, unvoiced, n/a
manner	fric.; stop, closure, flap, nasal, approx.; nasal flap, n/a
place	lab.; dent.; alveolar, pal.; vel.; glot.; lat.; rhotic, n/a
height	high, mid, low, lowhigh, midhigh, n/a
front	front, back, central, backfront, n/a
round	round, nonround, round-nonround, nonround-round, n/a
tense	tense, lax, n/a

Table 8. Phonetic attributes extracted by the MLPs in (Morris & Fosler-Lussier, 2006).

Linking knowledge from Brno University of Technology and the Georgia Institute of Technology, we get one of the best reported results on the phone TIMIT recognition task.

The authors, Siniscalchi, Schwarz and Lee, (Siniscalchi et al., 2007) report 79% for accuracy in the complete TIMIT test set. The proposed system is similar to that described by Schwarz, Matejka, and Cernocky in (Schwarz et al., 2006) and can be seen as a TANDEM architecture of MLPs ending in a HMM decoder. The left and right signal contexts are processed separately with windowing and DCT transform and each is applied to a different neural network. The outputs of these two neural networks feed another neural network. Finally, an HMM decoder is then used to turn these last neural network outputs, which are frame-based, into a signal that is segmented in terms of phones.

An extra knowledge-based module to rescore the lattices is included in (Siniscalchi et al., 2007). The lattice rescoring is done in two phases. In the first, the decoder generates a collection of decoding hypotheses. It is followed by a rescoring algorithm that reorders these hypotheses at the same time as it includes additional information. The additional information comes from a bank of speech attribute detectors which capture articulatory information, such as the manner and place of articulation. The bank of speech attribute detectors uses HMMs to map a segment of speech into one of the 15 broad classes, i.e. fricative, vowel, stop, nasal, semi-vowel, low, mid, high, labial, coronal, dental, velar, glottal, retroflex, and silence.

A log-likelihood ratio (LLR) at a frame level is taken as the measure of goodness-to-fit between the input and the output of each detector. A feed-forward ANN is then trained to produce phone scores for each set of LLR scores. These phone scores are then used in the lattice rescoring process changing the value of the arcs as a weight sum between the original values, with these last coming from the attribute detectors. The set of phones used in this system is the same as in (Rose & Momayyez, 2007).

In early 2009, Hifny and Renals (Hifny & Renals, 2009) presented a phonetic recognition system where the acoustic modulation is achieved by means of augmented conditional random fields. The results, using the TIMIT database, are very good. They reach 73.4% for accuracy using the core test set and 77% in another test set which includes the complete test set and the SA sentences.

A new automatic learning technique for speech recognition applications has recently been presented (Mohamed & Hinton, 2010). The authors, Mohamed and Hinton, apply Restricted Bolzmann Machines (RBMs) to phonetic recognition. Boltzmann machine is a type of stochastic recurrent neural network. As a real generative model, the Trajectory-HMM overcomes a major weakness of using HMMs in speech recognition, which is the conditional independence assumption between state outputs. With respect to the TIMIT database the authors observe that RBMs outperform a conventional HMM based system in 0.6% of PER. Regarding accuracy and using the core test set the result is 77.3%. A recent publication (Mohamed et al., 2011) reports the use of neural networks for acoustic modelling, in which multiple layers of features are generatively pre-trained. The outcome is, to the best of our knowledge, the highest TIMIT results reported so far in the core test set, 79.3% accuracy.

Although a fair comparison cannot be always made, Table 9 summarizes some of what we believe are the most important systems, considered as milestones in TIMIT phone recognition over the past twenty years. The systems differ considerably in terms of features, test material, phone set, acoustic modelling etc.; which make their comparison harder. A timeline survey, including the speech technology involved, the achieved rates and the test set used is then presented.

Year	System	Speech Technology	%Corr	%Acc	Test Set
1989	(Lee & Hon, 1989)	HMM	73.80	66.08	160 utterances (TID7)
1991	(Robinson & Fallside, 1991)	Recurrent Error Propagation Network	76.4 76.5	68.9 69.8	160 utterances (TID7) Complete Set
1992	(Young, 1992)	HMM	73.7	59.9	160 utterances randomly selected
1993	(Lamel & Gauvain, 1993)	Triphone Continuous HMMs	77.5	72.9	Complete Set
1994	(Robinson, 1994)	RNN	78.6 77.5	75.0 73.9	Complete Set Core Set
1998	(Halberstadt & Glass, 1998)	Heterogeneous input features. SUMMIT. Broad classes	-	75.6	Core Set
2003	(Reynolds & Antoniou, 2003)	MLP, Broad Classes	-	75.8	1152 utterances
2006	(Sha & Saul, 2006)	GMMs trained as SVMs	-	69.9	Complete Set
2006	(Schwarz et al., 2006)	TRAPs and temporal context division	-	78.52	Complete Set
2007	(Deng & Yu, 2007)	Hidden Trajectory Models	78.40	75.17	Core Set
2007	(Rose & Momayyez, 2007)	TDNN, phonological features HMM		72.2	Complete Set
2007	(Scanlon et al., 2007)	MLP/HMM	-	74.2	Complete Set
2007	ASAT, (Bromberg et al., 2007)	MLP/HMM	73.39	69.52	-
2007	(Siniscalchi et al., 2007)	TRAPs, temporal context division + lattice rescoring	-	79.04	Complete Set
2008	(Morris & Fosler-Lussier, 2006)	MLP/CRF	- 74.76	70.74 71.49	Core Set 944 utterances
2009	(Hifny & Renals, 2009)	Augmented CRFs	-	77.0	Complete Set + SA
2010	(Mohamed & Hinton, 2010)	Boltzmann Machines	-	77.3	Core set
2011	(Mohamed et al., 2011)	Monophone Deep Belief Networks	-	79.3	Core set

Table 9. Milestones in phone recognition using the TIMIT database. The percentages of correctness (%Corr) and accuracy (%Acc) are given.

4. Conclusions and discussion

This chapter focuses on how speech technology has been applied to phone recognition. It contains a holistic survey of the relevant research on the TIMIT phone recognition task, spanning the last two decades. This survey is intended to provide baseline results for the TIMIT phone recognition task and to outline the research paths followed, with varying success, so that it can be useful for researchers, professionals and engineers specialized in speech processing when considering future research directions.

The previous section described several approaches for phone recognition using the TIMIT database. Fig. 1 shows the chronology of the milestones in TIMIT phone recognition performance. Over the past 20 years the performance improved about 13%, mainly in the first 5 years of research. Improvement in the last 15 years has been very slight. Several approaches, covering different original technologies have been taken, none of them entirely solving the problem. It is hard to extrapolate future improvements from the graph in Fig. 1, but it appears that an upper bound of about 80% for accuracy will be hard to beat.

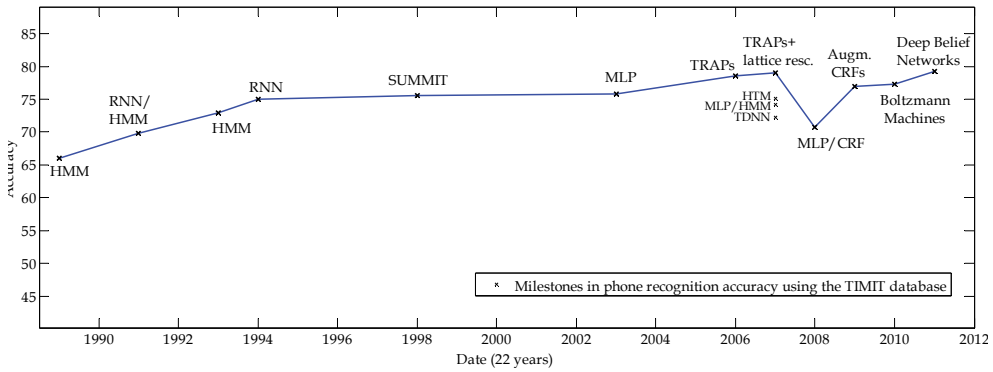

Fig. 1. Progress of the performance of TIMIT phone recognizer milestones.

Is there room for further improvement? Or does TIMIT database itself not allow it? The TIMIT hand-labelling was carefully done, and the labels have been implicitly accepted by the research community. Nevertheless, some authors (Keating et al., 1994; Räsänen et al., 2009) have pointed out issues related to TIMIT annotation. In (Keating et al., 1994) phonetic research on TIMIT annotations is described, drawing a parallel between standard and normative descriptions of American English. In spite of raising a question about the theoretical basis of the segmental transcriptions, the authors still found them useful. Another issue relates to label boundaries. In TIMIT 21.9% of all boundaries are closer than 40 ms to each other, (Räsänen et al., 2009). This may potentiate deletion errors, as a typical frame rate is 10ms, resulting in phones with less than 4 frames. Does this have an impact on the systems' performance, restricting room for improvement? Maybe, but we think that the long tradition of using the TIMIT test sets as a way of comparing new systems and approaches in exactly the same conditions will prevail.

Although the data in Fig. 1 indicate that there is limited room for improvement, new challenges must be taken up so as to uncover the full potential of speech technology. Until now, the main research issues rely on discriminative approaches and on the use of additional information, mainly wider feature temporal context, as well as speech attributes,

broad phonetic groups, landmarks, and lattice rescoring. The acoustic and phonetic information in the speech signal might already be fully exploited. One way to create a breakthrough in performance might be by adding syntactic (although language models are often used) or higher linguistic knowledge. Decoding the "meaning" of the words will probably help to improve word recognition, implying a top down approach or even avoiding classification at the phone level.

5. Acknowledgment

Carla Lopes would like to thank the Portuguese foundation, Fundação para a Ciência e a Tecnologia, for the PhD Grant (SFRH/BD/27966/2006).

6. References

Baker, J. M.; Deng, L.; Glass, J.; Khudanpur, S.; Lee, C.; Morgan, N. & O'Shaugnessy, D. (2009a). Research Developments and Directions in Speech Recognition and Understanding, Part 1. *IEEE Signal Processing Magazine*, vol. 26, no. 3, (May 2009), pp. 75-80, ISSN 1053-5888.

Baker, J. M.; Deng, L.; Glass, J.; Khudanpur, S.; Lee, C.; Morgan, N. & O'Shaugnessy, D. (2009b). Updated MINDS Report on Speech Recognition and Understanding, Part 2. *IEEE Signal Processing Magazine*, vol. 26, no. 4, (July 2009), pp. 78-85, ISSN 1053-5888.

Bromberg, I.; et al.. (2007). Detection-based ASR in the automatic speech attribute transcription project. *Proceedings of Interspeech2007*, ISSN 1990-9772, Belgium, August, 2007.

Deng, L. & Yu, D. (2007). Use of differential cepstra as acoustic features in hidden trajectory modelling for phonetic recognition. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* Hawaii, April, 2007.

Deng, L.; Yu, D. & Acero, A. (2005). A generative modeling framework for structured hidden speech dynamics. *Proceedings of NIPS Workshop on Advances in Structured Learning for Text and Speech Processing*, Canada, December 2005.

Dong , Y.; Deng, L. & Acero, A. (2006). A lattice search technique for a long-contextual-span hidden trajectory model of speech. *Speech Communication*, 48: pp: 1214-1226, ISSN 0167-6393.

Fujihara, H. & Goto, M. (2008). Three Techniques for Improving Automatic Synchronization between Music and Lyrics: Fricative Detection, Filler Model, and Novel Feature Vectors for Vocal Activity Detection. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* pp.69-72, USA, April 2008.

Furui, S. (2005). 50 Years of Progress in Speech and Speaker Recognition Research. *Proceedings of ECTI Transactions on Computer and Information Technology*, vol. 1, no. 2, 2005.

Garofolo, J.; Lamel, L.; Fisher, W.; Fiscus, J.; Pallett, D.; & Dahlgren, N. (1990). DARPA, TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. National Institute of Standards and Technology, 1990.

Gruhne, M.; Schmidt, K.; Dittmar, C. (2007). Phone recognition in popular music. *Proceedings of 8th International Conference on Music Information Retrieval*, Austria, September 2007.

Halberstadt, A. K. (1998). Heterogeneous Acoustic Measurements and Multiple Classifiers for Speech Recognition. *Ph.D. thesis*, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1998.

Halberstadt, A. & Glass, J. (1998). Heterogeneous measurements and multiple classifiers for speech recognition. *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, Australia, November 1998.

Hifny Y. & Renals S. (2009). Speech recognition using augmented conditional random fields. *IEEE Transactions on Audio, Speech & Language Processing*, vol. 17, no. 2, 2009, pp. 354–365, ISSN 1558-7916.

Hosom, J.-P. (2009). Speaker-independent phoneme alignment using transition-dependent states. *Speech Communication*, vol. 51, no. 4, 2009, pp. 352–368, ISSN 0167-6393.

Jiang, H. (2010). Discriminative training of HMMs for automatic speech recognition: A survey. *Computer Speech & Language*, Vol. 24, No. 4, October 2010, pp. 589-608, ISSN 0885-2308.

Junqua, J.-C. & Haton J.-P. (1996). Robustness in Automatic Speech Recognition: Fundamentals and Applications". Boston: *Kluwer Academic Publishers*. ISBN 0792396464.

Karsmakers, P.; Pelckmans, K.; Suykens, J.; Van Hamme, H. (2007). Fixedsize kernel logistic regression for phone classification. *Proceedings of Interspeech2007*, ISSN 1990-9772, Belgium, August, 2007.

Keating, P.A.; Byrd, D.; Flemming, E. & Todaka Y. (1994). Phonetic analyses of word and segment variation using the TIMIT corpus of American English, *Speech Communication*. 14, 131-142, ISSN 0167-6393.

Lamel L. & Gauvain J. L. (1993). High Performance Speaker Independent Phone Recognition using CDHMM. *Proceedings of Eurospeech*, Germany, September, 1993.

Lee, C.-H.; Clements, M. A.; Dusan, S.; Fosler-Lussier, E.; Johnson, K.;. Juang, B.-H and Rabiner, L. R. (2007). An Overview on Automatic Speech Attribute Transcription (ASAT). *Proceedings of Interspeech2007*, ISSN 1990-9772, Belgium, August, 2007.

Lee, K. & Hon, H. (1989). Speaker-independent phone recognition using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol.37 (11), November 1989, pp. 1642-1648, ISSN: 0096-3518.

Li, J. (2008). Soft Margin Estimation for Automatic Speech Recognition. *PhD thesis*, Georgia Institute of Technology, School of Electrical and Computer Engineering, 2008.

Matejka, P. (2009). Phonotactic and Acoustic Language Recognition. *PhD thesis*, Brno University of Technology, Faculty of Electrical Engineering and Communication, 2009.

Mohamed, A.; Dahl, G.; Hinton, G. (2011). Acoustic Modeling using Deep Belief Networks", *IEEE Transactions on Audio, Speech, and Language Processing*, 2011, ISSN 1558-7916.

Mohamed, A.-R.; Hinton, G. (2010). Phone recognition using Restricted Boltzmann Machines. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* USA, March 2010.

Morris, J. & Fosler-Lussier, E. (2008). Conditional Random Fields for Integrating Local Discriminative Classifiers. *IEEE Transactions on Audio, Speech, and Language Processing*, 16:3, March 2008, pp 617-628. , ISSN 1558-7916.

Morris, J. & Fosler-Lussier, E. (2007). Further experiments with detector-based conditional random fields in phonetic recognition. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* Hawaii, April 2007.

Morris, J. & Fosler-Lussier, E. (2006). Combining phonetic attributes using conditional random fields. *Proceedings of Interspeech2006*, USA, September 2006.

Ney, H. & Ortmanns, S. (2000). Progress in Dynamic Programming Search for LVCSR. *Proceedings of the IEEE, 88(8):1224-1240*, 2000.

Räsänen, O.; Laine, U. & Altosaar, T. (2009). An Improved Speech Segmentation Quality Measure: the R-value. *Proceedings of Interspeech2009*, U.K.; September 2009.

Reynolds, T.J. & Antoniou, C.A. (2003). Experiments in speech recognition using a modular MLP architecture for acoustic modelling. *Information Sciences*, vol 156, Issue 1-2, 2003, pp 39 – 54, ISSN 0020-0255.

Robinson T. & Fallside F. (1991). A Recurrent Error Propagation Network Speech Recognition System. *Computer Speech & Language*, 5:3, 1991, pp. 259-274, ISSN 0885-2308.

Robinson, T. (1994). An application of recurrent nets to phone probability estimation. *IEEE Transactions on Neural Networks*, vol. 5, no. 2, March 1994, ISSN 1045-9227.

Robinson, T.; Almeida, L.; Boite, J. M.; Bourlard, H.; Fallside, F.; Hochberg, M.; Kershaw, D.; Kohn, P.; Konig, Y.; Morgan, N.; Neto, J. P.; Renals, S.; Saerens, M.; & Wooters, C. (1993). A neural network based, speaker independent, large vocabulary, continuous speech recognition system: The wernicke project. *Proceeding of Eurospeech93*, Germany, September 1993.

Rose, R.; Momayyez, P. (2007). Integration Of Multiple Feature Sets For Reducing Ambiguity In ASR". *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2007 (ICASSP),* Hawaii, April 2007.

Scanlon, P.; Ellis, D. & Reilly, R. (2007). Using Broad Phonetic Group Experts for Improved Speech Recognition. *IEEE Transactions on Audio, Speech and Language Processing*, vol.15 (3) , pp 803-812, March 2007, ISSN 1558-7916.

Schwarz, P. (2008). Phone recognition based on long temporal context. PhD thesis, Brno University of Technology, Faculty of Information Technology, 2008.

Schwarz, P.; Matejka, P. & Cernocky, J. (2006). Hierarchical structures of neural networks for phone recognition. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2006 (ICASSP),* France, May 2006.

Sha, F. & Saul, L. (2006). Large margin Gaussian mixture modelling for phonetic classification and recognition. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2006 (ICASSP),* France, May 2006.

Siniscalchi, S. M.; Schwarz, P. & Lee, C.-H.; (2007). High-accuracy phone recognition by combining high-performance lattice generation and knowledge based rescoring. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2007 (ICASSP),* Hawaii, April 2007.

Young, S. J. (1992). The general use of tying in phone-based hmm speech recognizers. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 1992 (ICASSP),* USA, March 1992.

Young, S. J.; et al. (2006). The HTK book. Revised for HTK version 3.4, *Cambridge University Engineering Department*, Cambridge, December 2006.

Zue, V. & Seneff, S. (1996). Transcription and alignment of the TIMIT database. *In Hiroya Fujisaki (Ed.), Recent research toward advanced man-machine interface through spoken language*. Amsterdam: Elsevier, pp 464-447, 1996.

Zue, V.; Seneff, S. & Glass J. (1990). Speech database development at MIT: TIMIT and beyond. *Speech Communication*, Vol. 9, No. 4, pp. 351-356, ISSN 0167-6393.

Speech Technologies

Edited by Prof. Ivo Ipsic

ISBN 978-953-307-996-7
Hard cover, 432 pages
Publisher InTech
Published online 23, June, 2011
Published in print edition June, 2011

This book addresses different aspects of the research field and a wide range of topics in speech signal processing, speech recognition and language processing. The chapters are divided in three different sections: Speech Signal Modeling, Speech Recognition and Applications. The chapters in the first section cover some essential topics in speech signal processing used for building speech recognition as well as for speech synthesis systems: speech feature enhancement, speech feature vector dimensionality reduction, segmentation of speech frames into phonetic segments. The chapters of the second part cover speech recognition methods and techniques used to read speech from various speech databases and broadcast news recognition for English and non-English languages. The third section of the book presents various speech technology applications used for body conducted speech recognition, hearing impairment, multimodal interfaces and facial expression recognition.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821