AUTOMATIC AND LANGUAGE INDEPENDENT TRIPHONE TRAINING USING PHONETIC TABLES

Lorin Netsch and Alexis Bernard

DSP Solutions R&D Center, TEXAS INSTRUMENTS
12500 TI Blvd MS 8649, Dallas, TX 75243
{netsch,bernard}@ti.com

ABSTRACT

Training triphone acoustic models for speech recognition is time-consuming and requires important manual intervention. We present an alternative solution, performing *automatic* training by use of a pronunciation phonetic table which summarizes the articulatory characteristic of the target language. The method is able to train triphones for any language given an existing set of reference monophones in one or more languages by automatically performing the tasks of monophone seeding, triphone clustering and other training steps. The automatic nature of the training algorithm lends itself to parameter optimization, which can further improve recognition accuracy with respect to manually trained models. In a continuous digit recognition experiment, it is shown that automatically generated triphone models gave a 1.26% error rate, compared to a 2.30% error rate for its manual counterpart.

1. INTRODUCTION

Automatic speech recognition (ASR) is accomplished by determining the words that were most likely spoken, given a speech signal. This is done by comparing a set of parameters describing the speech signal with a set of trained acoustic model parameters. Hence, much effort is expended to produce acoustic models that provide the level of performance desired. The units of trained acoustic models may correspond to words, monophones, biphones or triphones. Triphones, which comprehend the prior and subsequent phone context of a given phone, typically outperform monophones, and are often the acoustic models of choice.

While triphones provide better performance, the number of triphones is often larger than the number of monophones by two orders of magnitude. Training thousands of triphones is complex and time-consuming. Some steps are machine intensive; others require a great deal of human intervention, which is error-prone. Such elements impact the cost and time to market associated with training acoustic triphone models for any new language.

Current manual acoustic training techniques are well known [1], and would have to be re-applied for any new target language, with careful human attention to adapt the training scheme to the characteristics of the new language.

Another method sometimes used to obtain acoustic models in a new target language is to adapt existing acoustic models in a reference language using a small database in the target language, at the cost of reduced recognition performance in the target language.

Previous research has dealt with the task of rendering automatic some aspects of acoustic training, such as [2] which automatically defines phonetic questions using intermediate clusters from a phoneme clustering algorithm. Others [3, 4] propose unsupervised training algorithms which use the technique of bootstrapping to minimize the amount of manual training or transcription time. In this paper, we propose a novel and simple framework which can be used to perform completely unsupervised and highly performing training in any language.

2. AUTOMATIC TRAINING USING PHONETIC TABLES

We propose a new automatic training technique which can operate on any language with virtually no human interaction and without sacrificing performance. The method developed only needs as input the following elements:

- Phonetic table(s), which characterize(s) the phones used in one or more *reference* language(s) with respect to their articulatory properties.
- A phonetic table, which characterizes the phones used in the *target* language with respect to their articulatory properties.
- A set of trained monophones for each of the *reference languages*.
- A database of sentences in the *target* language and its phonetic transcription of the database.

With those inputs, the proposed method completely and automatically takes care of all remaining manual steps, monophone seeding and triphone clustering, and machine intensive training steps involved in triphone acoustic training. The overall diagram of the designed method is presented in Figure 1.

In the proposed method, we reduce the amount of human intervention necessary to create acoustic models in a new target language to only describing the phonetic characteristics of the target language. First, the information about the target language summarized in its *phonetic table* is compared with that of reference languages for which we already possess trained acoustic models. Second, an intelligent algorithm creates seed acoustic models for the target language using the reference language acoustic models and the two phonetic tables. Furthermore, all the necessary human intervention steps associated with completely retraining the target language seed model are automatically performed by the computer by use of the phonetic table, as will be described.

Note that while our acoustic modelling uses hidden Markov models (HMM), operates on Mel-frequency cepstral coefficients (MFCC), and delivers trained triphones, the method can be extended to any type of speech modelling (dynamic time warping,

0-7803-8484-9/04/$20.00 ©2004 IEEE

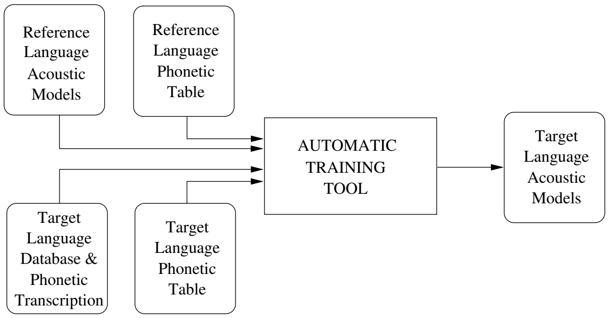
Reference Language Acoustic Models
Reference Language Phonetic Table
AUTOMATIC TRAINING TOOL
Target Language Acoustic Models
Target Language Database & Phonetic Transcription
Target Language Phonetic Table

Fig. 1. Block diagram of the proposed method performing automatic training in any language.

etc.), speech features (linear prediction cepstral coefficients, LPCC or perceptual linear prediction, PLP) and training units (monophones, biphones, triphones, etc.). The training method can also be applied to any pair of reference and target languages.

Finally, instead of using one reference language, the reference language phonetic table can consist of phones from several reference languages for which we already have trained acoustic models. This way, the pool of available phones is always increasing.

3. TYPICAL TRAINING PROCEDURE

For a better understanding of the scope and limitation of this method, the different steps typically required for acoustic triphone training are summarized. The usual first steps for acoustic triphone training in any language are given hereafter [1]. Many additional and refining steps can be applied at the end, such as augmentation of the number of mixtures, tying variances, separating male from female speech, performing vocabulary specific training, etc.

Monophone seeding: Monophone seeding constitutes the foundation of any training method. Subsequent training steps in any language of consideration are indeed based on the initial monophone seeds. Such monophone models can easily be estimated if one possesses a database that has been labeled and time marked all the way to the monophone level. This labeling and time marking requires extensive human intervention and thus is rarely performed.

Alternatively, seed monophones can be obtained through bootstrapping, which makes an estimate of the monophones using other already trained acoustic models depending on their acoustic similarities. While this technique is useful if the monophone similarities can be clearly estimated, it often requires a great deal of human interaction both to analyze which monophones are similar acoustically and to adapt topology of the reference model to fit with that of the target model.

If no other method is available, monophone seeding may use a simple ”flat start” method, whereby one initial model is constructed based on global statistics of the entire target training database. This model is duplicated to form the model for all monophones. This technique is rarely used for high-end speech recognition systems because it significantly impacts recognition performance.

Monophone training: Seed monophones are re-estimated using the entire target language database. Note that such re-estimation tries to maximize the likelihood of the training database given the speech models and that such local maximization operation is highly dependent on the initial (seeds) monophone models.

Monophone cloned into triphone: Seed triphones `a-b+c` are obtained by cloning the monophone `b`. An alternative method consists of creating seed triphones by performing forced alignment of the training data using monophones, generating the associated triphone context, and then training seed triphones using training data corresponding to the location of the triphones.

Triphone training: Each triphone is retrained using the entire target language training database.

Triphone clustering: The large number of triphones results in an excessive number of model parameters that must be trained, which requires extremely large training databases in order to successfully estimate the parameters. In order to reduce the number of parameters needed to represent the triphone models, after preliminary training of the triphone models, another procedure clusters the parameters. During clustering, parameters of similar triphones are linked together to obtain a joint and therefore more robust estimate of the clustered triphone parameters. The success of clustering is based on correctly identifying the parameters that are correlated with each other and should be grouped.

Existing methods of clustering triphone model parameters require significant human involvement. Such techniques can be either data driven or tree based. In the first case, triphones that tend to produce similar speech features are clustered. One limitation of data driven clustering is that it does not deal with triphones for which there are no examples in the training data. In the second case, a phonetic binary decision tree is built, with yes/no questions attached at each node. All triphones in the same leaf node are then clustered. With such a tree, any triphone in the language can be constructed, if the tree questions are based on articulatory features of phones. Before any tree building can take place, all of the possible phonetic questions must be manually determined depending on the specific set of phonemes characterizing the target language and their articulatory phonetic characteristic (e.g. voiced/unvoiced, place and manner of articulation, position of the tongue and jaw, strident, open jaw, round lips, long).

The disadvantage of direct application of these existing training techniques is time and cost associated with human intervention which needs to be repeated for each additional language. In addition, the resulting acoustic model sets are not optimized by selecting the best candidate from the large multitude of possible clustering candidates, resulting in degraded speech recognition performance and/or excessive model size.

Clustered triphone training: The clustered triphones are reestimated.

Subsequent training operations, such as increasing the number of Gaussian mixtures per state, separating male and female training or further tying HMM parameters to obtain model size reduction can always be applied.

4. THE PHONETIC TABLES

Table 1 gives an example of the phonetic table designed in order to describe the English language according to articulatory criterion. Table 1 is only one example of the concept of the proposed method.

Table 1 is subdivided into three classes of phonemes: vowel, consonants-semivowels and silences/closures.

Classification of phones in any language is determined according to the following articulatory or phonetic properties:

- Phone: There exist 43 phones in English.

- Class of phone: Vowel, diphthong, consonant, semi-vowel, or closure.
- Topology: Number of states in the HMM model.
- Length of phone: Short or long.
- Position of jaw: High, medium or low.
- Position of articulation: Front, central or back.
- Vowel type: A, E, I, O or U.
- Voicing: Voiced or unvoiced.
- Continuance: Continuant or non-continuant.
- Rounding of lips: Round or not.
- Tension in cheeks: Tense or lax.
- Manner of articulation: Stop, affricate, fricative, nasal, liquid, retroflex or glide.
- Point of articulation: Bilabial, labial, velar, alveolar, labiodental, alveopalatal or interdental.
- Stridency: Strident, non-strident, or unstrident.
- Zone of articulation: Anterior or non-anterior.
- Position of front of the tongue: Whether the consonant is coronal or not.
- Degree of muscular effort: Fortis, lenis, or neither.

For *vowel* sounds, the following classes apply, given in order according to Table 1: phone, topology, class, position of jaw, position of articulation, vowel type, voicing, continuance, length, tension and nasality.

For *consonant* sounds, the following classes apply: phone, topology, class, manner of articulation, position of articulation, point of articulation, voicing, continuance, muscular effort, position of the front of the tongue, zone of articulation and stridency.

VOWELS AND DIPHTHONGS											
/iy/	7	VOW	H	Fr	Iv	V	C	$\overline{\text{R}}$	Lo	LA	
/ih/	6	VOW	H	Fr	Iv	V	C	$\overline{\text{R}}$	Sh	TE	
/uw/	7	VOW	H	Ba	Uv	V	C	R	Lo	LA	
/uh/	6	VOW	H	Ba	Uv	V	C	R	Sh	TE	
/ah/	6	VOW	M	Ce	Av	V	C	$\overline{\text{R}}$	Sh	TE	
CONSONANTS AND SEMIVOWELS											
/d/	5	CON	STOP	Ce	ALV	V	$\overline{\text{C}}$	Len	Cor	$\overline{\text{A}}$	UNS
/k/	5	CON	STOP	Ba	VEL	$\overline{\text{V}}$	$\overline{\text{C}}$	For	$\overline{\text{Cor}}$	$\overline{\text{A}}$	UNS
/ch/	7	CON	AFF	Ce	ALVP	$\overline{\text{V}}$	$\overline{\text{C}}$	For	Cor	$\overline{\text{A}}$	STR
/jh/	5	CON	AFF	Ce	ALVP	V	$\overline{\text{C}}$	Len	Cor	$\overline{\text{A}}$	STR
/f/	7	CON	FRIC	Fr	LABD	$\overline{\text{V}}$	C	For	$\overline{\text{Cor}}$	A	NST
/v/	5	CON	FRIC	Fr	ALBD	V	C	Len	$\overline{\text{Cor}}$	A	NST
/m/	5	CON	NAS	Fr	BLB	V	C	Ufl	$\overline{\text{Cor}}$	A	UNS
/n/	5	CON	NAS	Ce	ALV	V	C	Ufl	Cor	$\overline{\text{A}}$	UNS
/l/	6	SEM	LIQ	Ce	ALV	V	$\overline{\text{C}}$	Ufl	Cor	$\overline{\text{A}}$	UNS
/el/	7	SEM	LIQ	Ce	ALV	V	$\overline{\text{C}}$	Ufl	Cor	$\overline{\text{A}}$	UNS
/y/	6	SEM	GLI	Ce	ALVP	V	$\overline{\text{C}}$	Ufl	Cor	$\overline{\text{A}}$	UNS
/r/	6	SEM	RET	Ba	ALV	V	$\overline{\text{C}}$	Ufl	$\overline{\text{Cor}}$	$\overline{\text{A}}$	UNS

Table 1. Portions of the phonetic table for the American English language. VOW=vowel, DIPH=diphtong, CONS=consonant, SEM=semivowel, CL=closure, SIL=silence, PAU=pause, H=high, M=medium, L=low, Fr=front, Ce=central, Ba=back, V=voiced, $\overline{\text{V}}$=unvoiced, C=continous, $\overline{\text{C}}$=non-continuous, R=round, $\overline{\text{R}}$=unround, Lo=long, Sh=short, LA=lax, TE=tense, STOP=stop, AFF=affricate, FRIC=fricative, NAS=nasal, LIQ=liquid, GLI=glide, RET=retroflex, BLB=bilabial, LAB=labial, ALV=alveolar, VEL=velar, ALVP=alveopalatal, LABD=labiodental, INTD=interdental, For=fortis, Len=lenis, Ufl=unfortlenis, Cor=coronal, $\overline{\text{cor}}$=non-coronal, A=Anterior, $\overline{\text{A}}$=non-anterior, UNS=unstrident, STR=strident, NST=non-strident

5. UTILIZATION OF THE PHONETIC TABLE FOR TRAINING PURPOSES

The proposed method allows for the steps of *monophone seeding* and *triphone clustering* to be automatically performed based on the phonetic table of the reference and target language.

5.1. Monophone seeding

The operation of obtaining monophone seeds for any training algorithm is often either imprecise (*e.g.* flat start) or very complex (*e.g.* requiring a manual phonetic transcription of all or parts of the database). In the proposed method, presented in Figure 2, we use the phonetic table of the reference language(s) (for instance English, Table 1) and a similar one for the target language to create seed monophone models for the target language in the following two steps:

Optimal match selection: Selecting for each phone in the target language, the phone in the reference language(s) that is the most similar in terms of articulatory characteristics. The assumption behind this operation is that regardless of the language, phones that are pronounced similarly will sound similar and have similar acoustic properties.

Change topology: Based on the topology (number of states) of the target language phone and its best match in the reference languages, the topology of the phone model in the reference language is modified to correspond to that of the reference language phone. The algorithm can take any pair of number of states (for the reference and target phone) and transforms the reference model using a weighted sum of the characteristics (means and variances) of each reference state in such a way that the target phone model represents a compressed or stretched version of the reference phone model.

5.2. Triphone clustering

As mentioned earlier, *tree based triphone clustering* is performed for two purposes: 1) clustering acoustically similar triphones together yields a more robust parameter estimate of those triphones, 2) tree based clustering allows for generation of models for triphones unseen in the training data.

Based on the same assumption as in Section 5.1, acoustic similarity can be implied from articulatory characteristics, and the phonetic table is re-used to build an acoustic decision tree adapted for each triphone sharing the same center monophone.

The triphone clustering operation, presented in Figure 3, is a three-tier process:

Universal questions: Generate a list of universal articulatory based questions. Being based on articulatory characteristics with categories shared among all languages, the list is language independent and can be referred to as the universal articulatory based question list. A non-exhaustive example of such universal question list is presented in Table 2.

Specific questions: Based on the phonetic table of the *tar-*

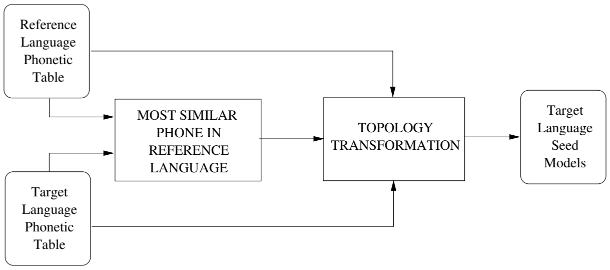
Reference Language Phonetic Table
MOST SIMILAR PHONE IN REFERENCE LANGUAGE
TOPOLOGY TRANSFORMATION
Target Language Seed Models
Target Language Phonetic Table

Fig. 2. Block diagram of the proposed method for monophone seeding using the phonetic table of the target and reference languages.

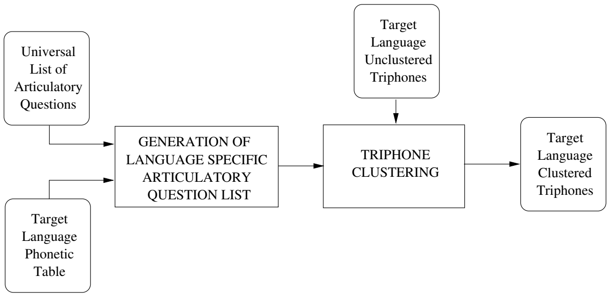
Universal List of Articulatory Questions
Target Language Unclustered Triphones
GENERATION OF LANGUAGE SPECIFIC ARTICULATORY QUESTION LIST
TRIPHONE CLUSTERING
Target Language Clustered Triphones
Target Language Phonetic Table

Fig. 3. Block diagram of the proposed method for triphone clustering using the phonetic table of the target language and the universal list of articulatory questions.

get language, transform each question of the universal question list into a language specific question which specifies exhaustively which phones conform to the articulatory characteristics asked by the question. Note that not all questions in the universal question list are applicable to the target language (for instance the nasal vowels do not exist in English but do exist in French). This is not an issue as such non-applicable questions will find no match in the target language and will simply be discarded.

Clustering: Using the target language specific question list, construct for all `a-b+c` triphones sharing the same center phone `b` an acoustic decision tree that maximizes the likelihood of observing the data by selecting at each node in the tree the question that most increases likelihood. This top-down approach is repeated until the increase in likelihood from further subdividing the triphones falls under a certain threshold.

Note finally that the phonetic table may also be used to automatize many training steps. For instance, the table is used in the following steps: create the monophone list, clone the monophones into triphones, increase the number of mixtures for each model and tie variances.

6. CONCLUSIONS

Under the assumptions that 1) articulatory characteristics are language independent since they are only a consequence of the human vocal apparatus and 2) sounds that are produced in a similar articulatory fashion will have similar acoustic properties, our solution combines all the information needed to train a new language into a phonetic table. Using the tables, we implement more accurate and completely unsupervised methods for monophone model seeding and triphone clustering. This, in turn, allows for the optimization of the phonetic models for best speech recognition performance and size tradeoff.

Question	**Definition**	**Question**	**Definition**
Boundary	SIL	Unvcd_Closure	CL & $\overline{V}$
Vcd_Closure	CL & V	Closure	CL
Front_stop	Fr & STOP	Central_stop	Ce & STOP
Back_stop	Ba & STOP	Affricate	AFF
Stop	STOP	Vcd_stop	V & STOP
Unvcd_stop	$\overline{V}$ & STOP	Nasal	NAS
Fric	FRIC, AFF	Fricative	FRIC
Voiced_Fric	(FRIC, AFF) & V	Vowel	VOW
Unvoiced_Fric	(FRIC, AFF) & $\overline{V}$	Front_vowel	Fr & VOW
Back_Fric	(FRIC, AFF) & Ba	Liquid	LIQ
Central_cons	Ce & CONS	Back_cons	Ba & CONS
Front	Fr	Central	Ce
Back	Ba	Fortis	For
Unvoiced_cons	$\overline{V}$ & CONS	Voiced_cons	V & CONS
Unvoiced	$\overline{V}$	Voiced	V
Bilabial	BLB	Labdental	LABD
Intdental	INTD	Alveolar	ALV
Alvpalatal	ALVP	Velar	VEL
Front_long	Fr & Lo	Mid_long	M & Lo
Back_long	Ba & Lo	High_long	H & Lo
Central_long	Ce & Lo	Low_long	L & Lo

Table 2. Portion of the universal list of articulatory questions

The advantage of an automatic script lies in the ability of launching a large training job without the need for supervision, manual editing of files or human interaction with the training algorithm. This represents an invaluable tool to 1) perform research in acoustic modelling, 2) improve present acoustic model performance by analyzing many different combinations of parameters or optimizing them, and 3) repeat that training task as many times as needed, for instance, considering many different languages, acoustic environments, or speech features.

Our method has been utilized to successfully train triphones in several languages, including American English, British English, German and Japanese. Repeatedly, it was seen that the proposed unsupervised training method outperformed the best handcrafted training method both in recognition accuracy and acoustic model size. For instance, when comparing continuous digit string recognition accuracy in American English, it was observed that automatically trained triphone acoustic models lead to a 1.26% word error rate, versus a rate of 2.30% for the manually trained models.

7. REFERENCES

[1] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (Version 3.0)*, July 2000.

[2] K. Beulen and H. Ney, “Automatic question generation for decision tree based state tying,” 1998, vol. 2, pp. 805–8.

[3] F. Wessel and H. Ney, “Unsupervised training of acoustic models for large vocabulary continuous speech recognition,” in *Automatic Speech Recognition and Understanding workshop*, 2001, pp. 307–10.

[4] T. Kemp and A. Waibel, “Unsupervised training of a speech recognizer: Recent experiments,” Sep. 1999, vol. 5, pp. 2725–28.