Phonetic-Search in a New Target Language Using Multi-Language Indexing and Phonetic-Mappings

Yossi Bar-Yosef, Ruth Aloni-Lavi, Irit Opher
NICE systems
Ra'anana, Israel
Yossi.Bar-Yosef;Ruth.Aloni-Lavi;Irit.Opher@nice.com

Ella Tetariy, Shiran Dudy, Vered Silber-Varod, Vered Aharonson, Ami Moyal
ACLP – Afeka Center for Language Processing
Afeka Academic College of Engineering
Tel Aviv, Israel
ellat;shirand;veredsv;vered;amim@afeka.ac.il

***Abstract*— The current paper considers methods for searching for spoken keywords in a new under-resourced target language using existing acoustic models of two other, highly resourced, source languages. The study addresses the framework of Phonetic-Search (PS) which is an extremely fast technique applied for spoken Keyword Spotting (KWS). To ensure accurate phonetic recognition in the indexing phase the phonetic model training requires substantial acoustic and linguistic resources, resulting in heavy and expensive operations. Furthermore, particular cases of under-resourced languages pose a real challenge for phonetic-search as the available linguistic resources are not enough for training acoustic models. In a preceding paper we introduced automatic learning of cross-language phonetic mappings from a single source language model set to a new target language phoneme set (i.e. providing one-to-one mappings). The current study extends the solution to performing the search over two phonetic lattices that were generated in the indexing phase by model sets of two different source languages. We provide comparative results of phonetic-search in Spanish and Dari as target languages, using American-English and Levantine Arabic as source languages. Results clearly indicate that fusing well two phonetic lattices that were acquired from different models can extend the phonetic coverage to improve phonetic search in a new target language.**

Keyword-spotting; phonetic-search; under-resourced languages; phonetic-mapping

I. INTRODUCTION

There is currently a growing demand for supporting new languages in Keyword Spotting (KWS) and other Automatic Speech Recognition (ASR) based applications. Supporting a new language requires a long and costly process of data collection and training of new acoustic models. Moreover, in some cases, and particularly for KWS in "exotic languages," sufficient training data is not available, altogether impeding the development of the application.

Since the 90's, research has focused on two different approaches for coping with this challenge. One approach uses phoneme sets and modeling from multiple languages to construct a global phone inventory suitable for a large group of languages [1,2], while the second either generates or adapts new acoustic models for new languages either by using manual or semi-automatic phoneme mappings [3], or by performing acoustic adaptation using a small corpus of the new language [4]. A recent study suggested using existing well-trained models from a few source languages for unsupervised transcription generation for training the under-resourced target language [5]. The methods of the latter two studies ([4] and [5]) involved using source language acoustic models for recognition in a target language, where some adaptation was applied after the initial mappings and alignments. However, all such attempts were aimed at Large Vocabulary Continuous Speech Recognition (LVCSR) or Language ID applications and not at KWS.

KWS based on Phonetic Search (PS) is an extremely fast technique that uses phonetic recognition in a pre-processing stage (regarded as the "indexing" phase) so that acoustic computations during the search phase can be avoided. The phonetic indexing aims at extracting the phonetic content of the speech, independently of possible required keywords. Moreover, PS systems usually use a phoneme-level Language-Model (LM) and not a word-level LM, and therefore are more flexible in describing the acoustic content.

In a preceding work [6], we introduced methods for applying PS in a new target language using existing acoustic models of another source language. We proposed a robust procedure for learning a statistical mapping between the new target phonemes and the existing phonetic models providing a "one-to-one" probabilistic mapping ("one-to-one" stands for *one source language* to *one target language*). PS is particularly suited for cross-language configurations for the following reasons: (1) the phonetic lattice represents the acoustic content of the speech; (2) the search is carried out through a series of "soft" decisions, depending on likelihoods into which mapping costs can be easily incorporated; (3) a word-level language model is not required.

The focus of this work is on improving phonetic search in a new target language given two phoneme lattices that were produced by two sets of acoustical models from two different source languages. This means that there is no intervention in the indexing phase (i.e. no change in the phoneme recognizers is performed), but only in the search phase. Even though we do not optimize the phoneme recognition toward the new target language, we still expect to gain additional information

This research is part of a grant (#45828) provided by the Chief Scientist of the Israeli Ministry of Commerce for developing Phonetic Search in New Languages Based on Cross-Language Transformations. The research was carried out as part of the Magneton program which encourages the transfer of knowledge from academic institutions to industrial companies – in this case ACLP – Afeka Center for Language Processing and Nice Systems Inc.

by looking at two separate recognition results resulting from the variations in the acoustic coverage of the different source languages.

In the current paper we address two approaches to perform KWS under the above conditions. First we examine a simple post-decision approach assuming that we are given KWS results from two separate "one-to-one" configurations, as described in [6] (for example "English-To-Spanish", and "Arabic-To-Spanish"). In the second approach, we fuse the two phoneme lattices (generated by "English" models and "Arabic" models) into a single unified lattice and then perform the search over a merged lattice. In our experiments, the second approach yielded superior and more robust results.

II. Background

A. Phonetic search

A keyword search over a recognized phoneme lattice is based on calculating the likelihood $p(\boldsymbol{O},\boldsymbol{R}|\boldsymbol{W})$, where we denote $\boldsymbol{O}=\{o_1,\ldots,o_T\}$ to be a series of T observation vectors, $\boldsymbol{R}=\{r_1,\ldots,r_S\}$ to be a recognized sequence of S phonemes, and $\boldsymbol{W}=\{w_1,\ldots,w_P\}$ as the searched word represented by a sequence of P phonemes. Namely, we need to compute the probability of observing $\boldsymbol{O}$ and recognizing $\boldsymbol{R}$, given that a particular keyword $\boldsymbol{W}$ was pronounced.

Using the simple Bayes' rule we obtain,

$$p(\boldsymbol{O},\boldsymbol{R}|\boldsymbol{W}) = p(\boldsymbol{O}|\boldsymbol{R},\boldsymbol{W})p(\boldsymbol{R}|\boldsymbol{W}), \quad (1)$$

and applying the Markov chain relation, $\boldsymbol{O}\leftarrow\boldsymbol{R}\leftarrow\boldsymbol{W}$, yields,

$$p(\boldsymbol{O},\boldsymbol{R}|\boldsymbol{W}) = p(\boldsymbol{O}|\boldsymbol{R})p(\boldsymbol{R}|\boldsymbol{W}). \quad (2)$$

Conveniently, the result in Eq. (2) is composed of two independent types of conditional probabilities. The left term, $p(\boldsymbol{O}|\boldsymbol{R})$ is the "acoustic" probability, and the right term, $p(\boldsymbol{R}|\boldsymbol{W})$ can be considered the "cross-phoneme (series)" probability. The major advantage of this solution is that the acoustic probabilities can be pre-calculated and stored as a phonetic lattice in the indexing phase, regardless of the searched keywords. The search process thus requires only the calculation of the "cross-phoneme" probabilities over the various paths in the recognized lattice. The search can be further simplified by the naive assumption that the cross-phoneme probabilities are context-independent. This leads to a factorial form of the likelihood computation such that

$$p(\boldsymbol{O},\boldsymbol{R}|\boldsymbol{W}) = p(\boldsymbol{O}|\boldsymbol{R})\prod_{i\in\boldsymbol{B}}p(r_i|w_i), \quad (3)$$

where $i\in B$ is the examined path, and noticing that w_i in the conditional probabilities, $p(r_i|w_i)$ can accommodate both insertion and deletion events. These phoneme-to-phoneme probabilities $p(r_i|w_i)$ are pre-defined in the system and are used by the search mechanism to compute the pattern matching scores. In practice, $p(\boldsymbol{R}|\boldsymbol{W})$ is computed through a dynamic-programming algorithm searching for the best matching path using $p(r_i|w_i)$ for the likelihood scoring.

Score Normalization: Most PS systems typically apply length normalization of the log-likelihood scores, in order to use a single decision threshold. The acoustic log-likelihood is normalized by the number of speech frames, and the phonetic matching log-likelihood is normalized by the number of phonemes of the searched keyword.

B. Cross-language phonetic mappings

In a preceding study [6] we investigated cross-language phonetic search from a single source language to a new target language. It was demonstrated how the modularity of PS (as reflected in Eq. (3)) can be leveraged to easily support a new target language if the cross language mapping used during the search phase are "sufficiently" accurate. In [6] it was assumed that acoustic model parameters of the source language remain fixed, and a suitable mapping $p(r|w)$ is used. Notice that $p(r|w)$ reflects the probability of recognizing a phonetic model r given that phoneme w of the target language was pronounced, and this mapping can be realized as a similarity or "confusion" matrix in the system, as illustrated in Figure 1.

Target (Spanish) \ Source (English)	AA	B	V	D	DH	EY	F
a	**0.18**	0	0.01	0.01	0	0	0
b	0.01	**0.28**	0.04	0.06	0	0	0.01
B	0.01	0.01	**0.18**	0.03	0	0	0.01
d	0	0.04	0.03	**0.35**	0.02	0	0.01
D	0.01	0.01	**0.06**	**0.08**	0.04	0	0.01
e	0	0.01	0.01	0.02	0	0	0.02
f	0.01	0	0.03	0	0	0	**0.54**

Fig. 1. Cross-language mapping illustration – each entry in the matrix indicates the likelihood p(r|w), which is the probability of recognizing source model r given that a target phoneme w was pronounced.

The work in [6] proposed methods to learn robust cross-language mappings, given a small amount of development data in the new target language. The basic idea is to start with an initial approximated mapping that is later used to perform series matching between the recognized best path (of phonetic models of the source language) and the phonetic sequence of a given keyword (from the target language). Using this mechanism, a statistical confusion matrix can be produced and used during the search.

The initial mapping may be obtained in two ways. In the first approach we use merely linguistic knowledge to formulate mapping rules that can be converted to a similarity matrix. This approach can be applicable when target development data is very limited. A second approach, involves an automatic acoustic distance calculation between low-order models of both the source and target languages. Assuming that we have small development data in the target language, it is possible to train low-order mono-phones of the new language and use them to compute approximated Kullback-Leibler (KL) distances between source and target language acoustic models. The acoustic distances can then be transformed to similarity measures (details are given in [6]) to obtain the required similarity matrix.

III. METHODS

In the current section we introduce a simple and effective method for fusing phonetic lattices that were generated by models of two different source languages in order to improve phonetic search in a new target language.

As mentioned, the focus of the current work only involves modifications in the search phase, while phonetic recognition using other source models remains as is. The goal is to extend the phonetic coverage that can be extracted along a search path without overly increasing the degrees of freedom in the search. In other words, the goal is to enable cross-language transitions between two phonetic lattices (that were generated by different models), while still restricting the search in order to avoid the inference of unrestrained paths that may eventually harm the overall accuracy.

Assume two independent cross-language configurations for a certain target language as described in section II-B, denoted by $\mathbf{A} \rightarrow \mathbf{T}$ and $\mathbf{B} \rightarrow \mathbf{T}$, where **A** and **B** symbolize the source languages and **T** symbolizes the new target language. For each configuration we are given a probabilistic mapping matrix $\boldsymbol{P}_{\mathbf{T|A}}$ and $\boldsymbol{P}_{\mathbf{T|B}}$ respectively, where each entry (t, s) in the matrix reflects a likelihood value of the form

$$\boldsymbol{P}_{\mathbf{A|T}}(t,s) = p(a_s|w_t)$$

$$\boldsymbol{P}_{\mathbf{B|T}}(t,s) = p(b_s|w_t),$$

where w_t is a target phoneme of **T**, a_s is a phonetic model of **A**, and b_s is a phonetic model of **B**. Notice that if we denote $|A|$ and $|B|$ as the size of the phoneme set in **A** and **B** respectively, and $|T|$ as the number of target phonemes in **T**, then it follows that $\boldsymbol{P}_{\mathbf{A|T}}$ is a $|T| \times |A|$ matrix, and $\boldsymbol{P}_{\mathbf{B|T}}$ is a $|T| \times |B|$ dimensional matrix[1].

Given the two configurations described above, we propose a simple method to construct a new multi-language configuration, $\mathbf{AB} \rightarrow \mathbf{T}$, as follows. First we define a new probabilistic mapping $\boldsymbol{P}_{\mathbf{AB|T}}$ by concatenating $\boldsymbol{P}_{\mathbf{A|T}}$ and $\boldsymbol{P}_{\mathbf{B|T}}$ (assuming the same phoneme order of $\boldsymbol{T}$ is inherent by the row order of both matrices), such that[2]

$$\boldsymbol{P}_{\mathbf{AB|T}} = [\, \boldsymbol{P}_{\mathbf{A|T}} \quad \boldsymbol{P}_{\mathbf{B|T}} \,]. \tag{4}$$

The next operation is implied per recording in the search process. Assuming that a recording was indexed by two phonetic lattices, $\boldsymbol{L_a}$ and $\boldsymbol{L_b}$, we produce a new lattice $\boldsymbol{L_{ab}}$ that allows cross-language transitions from one original lattice to another and vice versa in some constrained manner. If the phonetic lattice is expressed as a graph, where the nodes indicate time stamps and the arcs indicate the recognized phonemes, we add transition arcs between nodes of different source languages using a dedicated "pricing" rule. Noting that $|\Delta T_{ij}|$ represents the time gap between node i of lattice $\boldsymbol{L_a}$ and node j of lattice $\boldsymbol{L_b}$, we then apply a symmetric bi-directional transition between the two nodes with a log-likelihood cost that is given by

$$\xi_{ij} = \log[\, P_{cross}(\Delta T_{ij})] = -\epsilon_0 - \epsilon_1 |\Delta T_{ij}|, \tag{5}$$

where ϵ_0 and ϵ_1 are positive constants. Cross-lattice connections are illustrated in Figure 2.

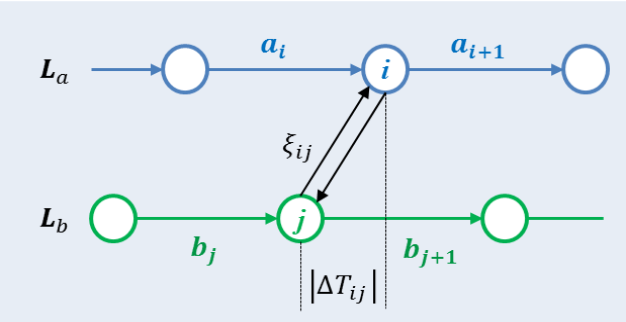

Fig. 2. Cross-lattice bi-directional transition between node i of lattice $\boldsymbol{L_a}$ and node j of lattice $\boldsymbol{L_b}$ with a symmetric transition cost, ξ_{ij}.

The cost rule in Eq. (5) penalizes the cross-language transitions but inserts some flexibility in time mismatches. Even when the time gap equals zero, it is necessary to set a small penalty, realized by $\epsilon_0 > 0$, to prevent loopbacks in the search. The second constant in (5), ϵ_1, controls the time difference penalty and can be calibrated to optimize search results. In practice, ΔT_{ij} is often quantized to frame-step units that typically relate to 10 millisecond time-steps. In our experiments we have observed that the transition cost should be significantly magnified within few frame steps, roughly between 5 to 10 frames, in order to constrain the search and avoid unreasonable paths.

Implementation issue: In order to reduce the computational cost during the search, it is preferable to prune arcs in the graph that entail very low transition probabilities. Through empirical experiments we have seen that a reduced form of lattice fusion can be adopted to save search computations. Apparently, it is sufficient to connect cross-language nodes within a time-gap of 30 milliseconds (namely up to 3-frame distance) with a minimal transition cost of $\xi_{ij} = -\epsilon_0$ (where $\epsilon_0 = 0.001$ and obviously ϵ_1 is set to zero). Eventually, this "economical" approach led to an almost negligible decrease in accuracy.

IV. EXPERIMENTS

This section reports on experiments held for two target languages, Spanish and Dari, given phoneme lattices indexed by the original models of two source languages, English and Arabic. Hence, in our experiments it was assumed that we have the following pre-trained one-to-one configurations, **En** → **Sp** (English-To-Spanish), and **Ar** → **Sp** (Arabic-To-Spanish) for Spanish; And for Dari, **En** → **Da** and **Ar** → **Da**, accordingly.

[1,2] To be more precise, the phonetic mapping matrices contain the "deletion event" in them, such that any referenced mapping matrix essentially includes an additional "target deletion" row, and an additional "source deletion" column.

Five corpora were used in the reported evaluations: English models where trained from the Wall Street Journal portion of Macrophone [7] that contains a collection of read sentences; Arabic models where trained using Levantine Arabic Conversational Telephone Speech [8] and Fisher Levantine Arabic Conversational Telephone Speech [9]; Spanish tests were performed on a portion of Spanish SpeechDat(II) FDB-4000 [10]; and Dari tests were performed using a portion of DAR_ASR001 from Appen.

Acoustic models were trained for both English and Arabic using the HTK toolkit [11]. An MFCC based, 39-dimensional, feature vector was used (13 Mel-Frequency Cepstral Coefficients, with the first and second derivatives), calculated over 25-millisecond frames with a 10 millisecond step. We used tri-phone modeling with HMMs containing 3 emitting states, each state's output probability was modeled by a mixture of 16 diagonal-covariance Gaussians. The search was performed on a list of keywords containing three or more syllables. The development set for estimating the confusion matrices included another hour of speech in the target language. Phoneme recognition was performed using HTK.

In order to evaluate the contribution of our suggested lattice fusion method we compared it to an additional post-decision stage that combines results that were independently generated by two corresponding one-to-one configurations. In the post-decision stage we tested several approaches to combine results. The approach referenced in this paper yielded the best keyword spotting performance that combined the results of **En → Sp** and **Ar → Sp**. In this quite simple approach, all keyword spotting results, from both configurations, are pooled together with additional score normalization that is *source-language*-dependent. The normalization, regarded as "Z-normalization" (Znorm), is performed for each cross-language configuration such that the normalized score is given by:

$$z = \frac{s-\mu}{\sigma},$$

where s is the raw score, and μ and σ are the mean and standard deviation of true-detection scores, computed over a small development set (less than half an hour).

In the following figures we show comparative results for different multi-language configurations with Spanish and Dari as target languages. In the figures, "**En**" and "**Ar**" correspond to English and Arabic source languages with the original one-to-one setting, and "**En+Ar**" relates to their combination. As mentioned, we examined a referenced post-decision technique that appears in the legend as "**En+Ar: union + Znorm**", and compared it to our suggested method for lattice fusion denoted as "**En+Ar: lattice fusion**". Addressing the results of Spanish, in Figure 3 it is observed that the simple post-decision approach with score normalization can boost the performance above the single source configurations. In addition, it is clearly observed that the proposed lattice fusion method provides significantly better results.

The Dari experiments posed a different situation where we have one cross-language configuration (**Ar → Da**) that is substantially superior to the other (**En → Da**). This case raises a serious difficulty in exploiting the weaker system to improve the performance of the better one using a post-decision approach.

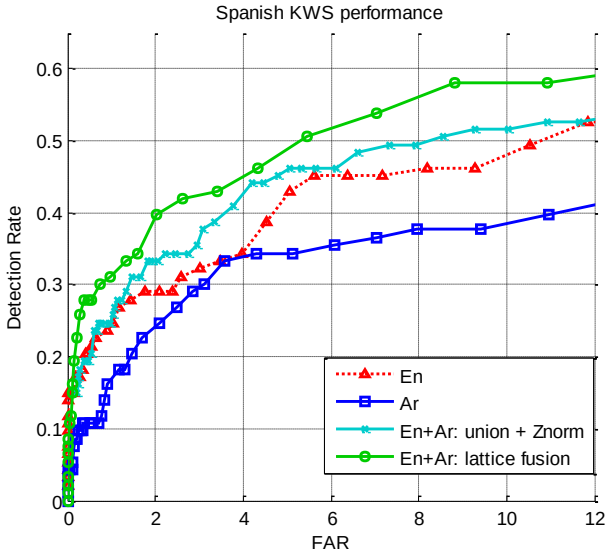

Fig. 3. KWS results for Spanish as a new target language with different multi-language configurations.

As shown in Figure 4, the Znorm scoring normalization led to degradation in accuracy, due to the fact that it upgraded the scores of the weak configuration, and thus inserted more false detections to the decision. Unlike the post-decision mechanism, the lattice fusion method provided a modest (but obvious) improvement that exceeded the performance of the **Ar → Da** system.

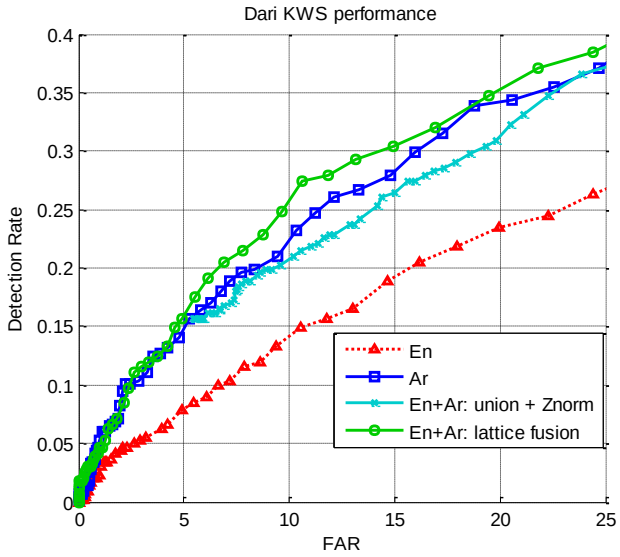

Fig. 4. KWS results for Dari as a new target language with different multi-language configurations.

To our understanding the robustness of the suggested fusion method lies in the probabilistic approach of the search

that is dependent on the mapping matrix, $\boldsymbol{P}_{\mathbf{AB|T}}$ (in Eq. (4)). In the Dari case for example, the search path will make a transition to an English route only when the phonetic mapping likelihood is high enough compared to other Arabic options, and thus in most cases only strong (in a probabilistic sense) English-Dari matches could affect the results, while the others are essentially neglected.

V. CONCLUSIONS

The paper has presented a lattice fusion approach for applying phonetic search in a new target language given phonetic models of two different source languages. Having two cross-language configurations with proper probabilistic mapping matrices (in each configuration a single source language is mapped to the new target language), we propose a simple implementation of a unified search by fusing the two related phonetic lattices and using a unified multi-language mapping matrix. The suggested approach adds flexibility to the search by allowing transitions between original lattices such that the phonetic content and context can be enriched in a single path. When done in a constrained manner, as described in the paper, the lattice fusion approach led to significant improvements in empirical experiments that were held. Under more difficult conditions, where one of the cross-language configurations is considerably weaker than the other (i.e. "English-To-Dari" compared to "Arabic-To-Dari"), the suggested method was still robustly able to exploit additional knowledge and exceed the performance of the stronger configuration (i.e. "Arabic-To-Dari").

As the described method is relatively simple and quite generic, it can be easily scaled-up to multiple lattice fusion of several different source languages.

REFERENCES

[1] T. Schultz and A. Waibel, "Fast bootstrapping of LVCSR systems with multilingual phoneme sets," in: *Proc. Eurospeech*, pp. 371-374, Rhodes, 1997.

[2] T. Schultz, "Globalphone: A multiligual speech and text database developed at Karlsruhe University" *ICSLP*, 2002.

[3] B. Wheatley et al., "An evaluation of cross-language adaptation for rapid HMM development in a new language," *ICASSP-94*, 1994.

[4] P. Fung, C.Y. Ma and W. K. Liu, "MAP-based cross-language adptation augmented by linguistic knowledge: from English to Chinese" *Eurospeech*, 1999.

[5] N. T. Vu, F. Kraus and T. Schultz, "Rapid building of an ASR system for under-resourced languages based on multilingual unsupervised training", *Interspeech*, 2011.

[6] Y. Bar-Yosef, R. Aloni-Lavi, I. Opher, N. Lotner, E. Tetariy, V. Silber-Varod, V. Aharonson and A. Moyal, "Automatic Learning of Phonetic Mappings for Cross-Language Phonetic-Search in Keyword Spotting", *IEEE 27th Convention of Electrical and Electronics Engineers in Israel*,, vol. 1, no. 5, pp.14-17, Nov. 2012.

[7] J. Bernstein, K. Taussig, and J. Godfrey, "MACROPHONE", *LDC*. Philadelphia, USA, 1994.

[8] Appen Pty Ltd, "Levantine Arabic Conversational Telephone Speech," *LDC*, Philadelphia, USA, 1994.

[9] M. Maamouri et al. "Fisher Levantine Arabic Conversational Telephone Speech", *LDC*, Philadelphia, USA, 2007.

[10] A. Moreno and J. A. Fonnolosa, "Spanish SpeechDat(II) FDB-4000 (ELRA-S0102)," *ELRA*, 2001.

[11] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, " The HTK Book, " HTK Version 3.0, Microsoft Corporation, July 2000.