Cross-Language Evaluation of Voice-To-Phoneme Conversions for Voice-Tag Application in Embedded Platforms

Yan Ming Cheng, Changxue Ma and Lynette Melnar

Human Interaction Research, Motorola Labs,
1925 Algonquin Road, Schaumburg, IL 60196, USA

Abstract

Previously, we proposed two voice-to-phoneme conversion algorithms for speaker-independent voice-tag creation specifically targeted at applications on embedded platforms, an environment sensitive to CPU and memory resource consumption [1]. These two algorithms (batch mode and sequential) were applied in a same-language context, i.e., both acoustic model training and voice-tag creation and application were performed on the same language.

In this paper, we investigate the *cross-language* application of these two voice-to-phoneme conversion algorithms, where the acoustic models are trained on a particular source language while the voice-tags are created and applied on a different target language. Here, both algorithms create phonetic representations of a voice-tag of a target language based on the speaker-independent acoustic models of a distinct source language. Our experiments show that recognition performances of these voice-tags vary depending on the source-target language pair, with the variation reflecting the predicted phonological similarity between the source and target languages. Among the most similar languages, performance nears that of the native-trained models and surpasses the native reference baseline.

Index Terms: voice-to-phoneme, voice-tag, speech recognition, embedded platform, cross-language, multilingual,

1. Introduction

A voice-tag (or name-tag) application converts human speech samples into an abstract representation which is then employed to recognize (or classify) speech in subsequent uses. The voice-tag application is the first widely deployed speech recognition application that uses technologies like DTW (Dynamic Time Warping) or HMMs in embedded platforms such as in mobile devices. Today, speaker-independent and phoneme HMM-based speech recognizers are also being included in mobile devices and voice-tag technologies are mature enough to leverage the existing computational resources and algorithms from the speaker-independent speech recognizer for further efficiency. For these reasons, the batch-mode and sequential voice-to-phoneme conversion algorithms were proposed in [1]. Both technologies create phonetic abstractions to represent a tag based on voice-to-phoneme conversions. The speaker-independent speech recognizer performs voice-tag recognition in the same way as any task performed by a phoneme-based speech recognizer. The performances of these algorithms have been shown to match or surpass that of the voice-tag

transcriptions made by expert phoneticians in a same-language environment.

One important advantage of previous voice-tag approaches, such as an HMM-based voice-tag technology, is their independence of existing language resources. With the globalization of mobile devices, this feature is imperative as it allows speaker-dependent speech recognition for resource-poor languages and dialects. Therefore, a legitimate concern confronting the proposed voice-tag approach is whether or not it can successfully leverage the speech resources from a resource-sufficient source language to recognize a target language for which little or no speech data is assumed. Several studies have in fact addressed the effectiveness of the cross-language application of phoneme acoustic models in speaker-independent speech recognition (see [2][3] [4][5]). In this paper, we attempt to demonstrate the cross-language effectiveness of the proposed voice-to-phoneme conversion algorithms in speaker-independent voice-tag applications for embedded platforms. In the next section, we briefly review the voice-to-phoneme conversion algorithms. In section 3, we describe the cross-language voice-tag experiments and provide the results in comparison with that of voice-tag applications in a same-language context. Finally, we share some concluding remarks in section 4.

2. Voice-to-phoneme conversion algorithms

In [1], two voice-to-phoneme conversion algorithms are proposed, namely the batch-mode and sequential voice-tag creations. They differ in how the voice-tag example utterances (hereafter example) are enrolled. Batch-mode conversion requires that all examples, usually two or three, are enrolled simultaneously and reaches its peak performance once enrollment is completed. Because performance improves as the number of enrollment examples increase, the enrollment process may lead to user frustration or even rejection. Sequential conversion, on the other hand, requires only one enrollment example per voice-tag, though successfully-recognized subsequent examples are used to update the voice-tag unobtrusively in the background. The obvious attraction of this algorithm is that the inconvenience of the enrollment process is minimized while performance is maximized, as there is no predefined limit on the number of examples per voice-tag. The less appealing feature of this strategy is that voice-tag recognition may be initiated with a lower performance.

Algorithmically, the two conversions differ in how to combine examples. The batch conversion combines examples at the feature level while the sequential conversion does so at the
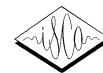
hypothesis level. Before we describe the algorithms, let us consider some of their common assumptions. There are M examples, $\mathbf{X_m}$ ($\mathbf{m} \in [1, \mathbf{M}]$), available to a voice-tag. $\mathbf{X_m}$ is a sequence of feature vectors corresponding to an example. Given the scope of our current discussion, we will not distinguish between a sequence of feature vectors and an example in the remaining part of this paper. The objective here is to find N phonetic strings, $\mathbf{P_n}$ ($\mathbf{n} \in [1, \mathbf{N}]$), following an optimization criterion.

2.1. Batch-mode voice-tag creation

The principle idea of batch-mode creation is to use a feature-based combination collapsing M examples into a single "average" utterance. The expectation is that this "average" utterance will preserve what is common in all of the constituent examples while neutralizing their peculiarities. Dynamic Time Warping (DTW) is used as the combination algorithm. Given two examples, $\mathbf{X_i}$ and $\mathbf{X_j}$ ($\mathbf{i} \neq \mathbf{j}$ and $\mathbf{i, j} \in [1, \mathbf{M}]$), a trellis can be formed with $\mathbf{X_i}$ and $\mathbf{X_j}$ being horizontal and vertical axes, respectively. Using a Euclidean distance and DTW algorithm, the best path can be derived, where "best path" is defined as the lowest accumulative distance from the lower-left corner to the upper-right corner of the trellis. A new example $\mathbf{X_{i,j}}$ can be formed along the best path of the trellis, $\mathbf{X_{i,j}} = \mathbf{X_i} \oplus \mathbf{X_j}$, where $\oplus$ is denoted as the DTW operator. The length of the new utterance is the length of the best path.

Let $\mathbf{X_{i,j}} = \{\mathbf{x_{i,j}}(0),..\mathbf{x_{i,j}}(\mathbf{t})..\mathbf{x_{i,j}}(\mathbf{L_{i,j}}-1)\}$,

$\mathbf{X_i} = \{\mathbf{x_i}(0),..\mathbf{x_i}(\varsigma)..\mathbf{x_i}(\mathbf{L_i}-1)\}$ and

$\mathbf{X_j} = \{\mathbf{x_j}(0),..\mathbf{x_j}(\tau)..\mathbf{x_j}(\mathbf{L_j}-1)\}$,

where $\mathbf{t}, \varsigma, \tau$ are frame indices. We define

$$\mathbf{x_{i,j}}(\mathbf{t}) = \frac{\mathbf{x_i}(\varsigma) + \mathbf{x_j}(\tau)}{2},$$ where $\mathbf{t}$ is the position on the best path

aligned to the ς-th frame of $\mathbf{X_i}$ and the τ-th frame of $\mathbf{X_j}$ according to the DTW algorithm. Applying repeatedly DTW operator M-1 times, we have the feature combination for M examples: $\mathbf{X_{1,2,3,...M}} = (\cdots((\mathbf{X_1} \oplus \mathbf{X_2}) \oplus \mathbf{X_3}) \cdots \oplus \mathbf{X_M})$. By using the "average" utterance and a simple phonetic decoder (see below), one can obtain **N** phonetic strings that serve as the abstract representation of the voice-tag. The phonetic decoder is usually a speech search engine constrained by a looped phoneme-grammar which is capable of delivering **N** best phonetic strings according to certain optimization criteria. (See [6] and [7] for a discussion of one such phonetic decoder.) Figure 1 depicts the batch-mode voice-creation system.
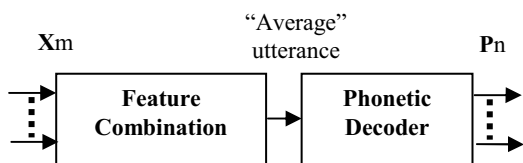

Figure 1: *Batch-mode voice-tag creation system*

2.2. Sequential voice-tag creation

Sequential voice-tag creation is based on the hypothesis combination of the outputs of a phonetic decoder of **M** examples. It is attractive for several reasons. First, only one example per voice-tag is required to create N initial seed phonetic strings, $\mathbf{P_n}$, using a phonetic decoder. The phonetic decoder is the same as described in the previous subsection. If good phonetic coverage (that is, good phonetic robustness of the trained HMMs) is exhibited by the phonetic decoder, with initial seed phonetic strings the recognition performance of voice-tags is usually acceptable, though not maximized. Each time a voice-tag is successfully utilized (i.e. a positive confirmation of the speech recognition result is detected and the corresponding action is implemented - for example, the call is made), the utterance is reused as another example to produce additional N phonetic strings to update the seed phonetic strings of the voice-tag through performing hypothesis combination. This update can be performed repeatedly until a maximum performance is reached. Figure 2 sketches this system.
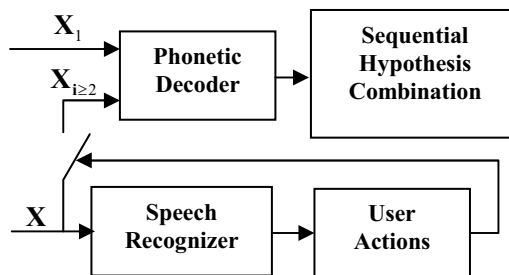

Figure 2: *Sequential combination of hypothetical results of a phonetic decoder*

The objective of this method is to discover a sequential hypothesis combination algorithm that leads to maximum performance. As detailed in **Error! Reference source not found.**, we use a hypothesis combination based on a consensus hierarchy displayed in the best phonetic strings of examples. The consensus hierarchy is expressed numerically in a phoneme n-gram histogram (typically a monogram or bigram is used). The sequential hypothesis combination algorithm is provided below:

- **Enrollment (or initialization)**: Use one example per voice-tag to create N phonetic strings via a phonetic decoder as the current voice-tag; use the best phonetic string to create the phoneme n-gram histogram for the voice-tag.
- **Step 1**: Given a new example of a voice-tag, create N new phonetic strings (via the phonetic decoder); update the phoneme n-gram histogram of the voice-tag with the best phonetic string of the new example.
- **Step 2**: Estimate a phoneme n-gram histogram per each phonetic string for N current and N new phonetic strings of the voice-tag.
- **Step 3**: Compare the phoneme n-gram histogram of the voice-tag with that of each phonetic string using a distance metric, such as divergence measure; select N phonetic strings, the histograms of which are closest to the histogram of the voice-tag histogram, as the updated voice-tag representation.

- **Step 4**: repeat steps 1-3 if a new example is available.

3. Experiments

3.1. Evaluation strategy and databases

The phonetic decoder we use in our experiments is MLite++, a Motorola proprietary HMM-based ASR search engine for embedded platforms, and a phoneme loop grammar. We use the ETSI advanced front-end standard for distributed speech recognition to generate a feature vector of 39 dimensions per frame. Context-dependent (CD) sub-word and speaker-independent HMMs are used for both the phonetic decoder and voice-tag speech recognition search engine.

In practice, evaluating cross-language performance is complex and poses distinct challenges to same-language performance evaluation. In general, cross-language evaluation can be approached by two principle strategies. The first strategy creates voice-tags in several languages by using language resources, such as HMMs and a looped phoneme grammar, from a single source language. The weakness of this strategy is that it is difficult to normalize the linguistic and acoustic differences across languages, a necessary step in creating an evaluation database. The second strategy creates voice-tags in a single language by using language resources from several distinct source languages. The weakness of this strategy is that language resources differ significantly and it cannot be expected that each source language will be trained with the same amount and type of data. Because we can compare our training data in terms of quantity and type, we opted to pursue the second strategy for the cross-language experiments presented here.

We select seven languages as source languages: British English (en-GB), German (de-DE), French (fr-FR), Latin American Spanish (es--LatAm), Brazilian Portuguese (pt-BR), Mandarin (zh-CN-Mand) and Japanese (ja-JP). For each of the source languages, we have sufficient data and linguistic coverage to train generic CD HMMs. The phoneme loop grammar of each source language is constructed from the phoneme set of that language.

Because we previously tested in [1] the performances of sequential and batch mode voice-to-phoneme conversion algorithms for speaker-independent voice-tag creation in an American English-only environment, and thus have these results for comparison, American English is chosen as the target language in the following cross-language experiments. The database selected for this evaluation is a Motorola-internal name database which contains a mixture of both landline and wireless calls. The database consists of spoken names of variable length and is divided into voice-tag creation and evaluation sets. The creation set has 85 name entries corresponding to 85 voice-tags, and each name entry comprises three examples spoken by a single speaker in different sessions. Thus the creation set is speaker-dependent. The purpose of designing a speaker-dependent creation set is that we expect that any given voice-tag will be created by a single user in real applications and not by multiple users. The evaluation set

contains 684 utterances of the 85 name entries. Most speakers of a name entry in the evaluation set are different from the speaker of the same name entry in the creation set, though some speakers are the same for both sets. In general, then, our evaluation set is speaker-independent.

3.2. Experiment results

In [1], we reported the sequential and batch-mode performance results in an American English-only environment: 95.2% and 91.23%, respectively. These performances are compared to the manual transcription baseline of 92.69%.

In the present investigation, the individual cross-language voice-tag recognition performances are compared to both the same-language results and to each other. To do the latter, a phonological similarity study is conducted between the target language (American English) and each of the selected evaluation languages, the prediction being that cross-language performance would correlate to the relative phonological similarity of the source languages to the target language. We use a pronunciation dictionary as each language's phonological description in order to ensure task independence, and because each language's pronunciations are transcribed in a language-independent notation system (similar to the International Phonetic Alphabet), cross-language comparison is possible [8]. Phoneme-bigram (biphoneme) probabilities collected from each dictionary are used as the numeric expression of the phonological characteristics of the corresponding language. The distance between the biphoneme probabilities of each source language and that of the target language is then measured. This metric thus explicitly provides a biphoneme inventory and phonotactic sequence importance. It also implicitly incorporates phoneme inventory and phonological complexity information. Using this method, the distance score is an objective indication of phonological similarity in the source-target language pair, where the smaller the distance value between the languages, the more similar the pair (see [2] for an in-depth discussion of this biphoneme distribution distance).

The languages that we use in these evaluations are from four language groups defined by genetic relation: (i) Germanic: en-US, en-GB, and de-DE; (ii) Romance: fr-FR, pt-BR, es--LatAm; (iii) Sinitic: zh-CN-Mand and (iv) Japonic: ja-JP. In general, it is expected that closely related languages and contact languages (languages spoken by people in close contact with speakers of the target language [9]), will exhibit greatest phonological similarity. The distance scores relative to American English are as follows, in order of increasing distance: en-GB, 0.61; de-DE, 1.46; fr-FR, 1.81; pt-BR, 1.82; zh-CN-Mand, 1.85; ja-JP, 1.90 and es--LatAm, 1.92. Note that the Germanic languages are measured to be the most similar to American English. In particular, the British dialect of English is least distant to American English, and German, the only other Germanic language in the evaluation set, is next. German is followed by French in phonological distance, and French and English are languages with centuries of close contact and linguistic exchange.

This preliminary study thus both substantiates in a quantitative way linguistic phonological similarity assumptions and provides

a reference from which to evaluate our results. Based on this study, it is our expectation that cross-language voice-tag application performance will be degraded relative to the voice-tag application performance in the same-language setting, and that the severity of the degradation will be a function of phonological similarity.

Table shows the cross-language voice-tag application performances of the sequential and batch mode voice-to-phoneme conversion algorithms, where the acoustic models are trained on the seven evaluation languages while the voice-tags are created and applied on American English, a distinct target language. For reference, we also include the American English HMM performance as a baseline.

Table 1: *Word Accuracies of voice-tag recognition with batch and sequential voice-tag creations in cross-language experiments.*

Sources	Word Acc. on Target Language		Distance
	Voice-Tag Creations		
	Sequential	Batch	
en-US (baseline)	95.2%	91.23%	0
en-GB	91.37%	87.13%	0.61
de-DE	90.50%	86.99%	1.46
fr-FR	89.91%	85.09%	1.81
pt-BR	82.75%	74.42%	1.82
zh-CN-Mand	92.11%	84.94%	1.85
ja-JP	78.07%	67.69%	1.90
es--LatAm	89.62%	83.33%	1.92

Apart from the exceptional performance of Mandarin using the sequential phoneme conversion algorithm, the performances generally adhere to the target-source language pair similarity scores identified above. Voice-tag recognition with British English-trained HMMs achieve a word accuracy of 91.37% while recognition with German-trained HMMs realize 90.5%. The higher-than-expected performance rate of Mandarin may be because the resources used to train Mandarin models are embedded with a significant amount of English material (English loan words, for example), reflecting a modern reality of language use in China.

The cross-language evaluations show significant performance differences between the two voice-creation algorithms across all of the evaluated languages. The differences are in accordance with our observation in the same-language evaluation. Although there are degradations, the performances of sequential voice-tag creation with HMMs trained on the languages most phonologically similar to American English are very close to the reference performance (92.69%) where the phonetic strings of voice-tags were transcribed manually by an expert.

4. Conclusions

We demonstrated that the voice-to-phoneme conversion algorithms proposed earlier for a same-language environment are also applicable in a cross-language setting, where HMMs trained on a source language are used in voice-tag creation and recognition of a distinct target language in an embedded platform. We used a distance metric to show that performance results associated with HMMs trained on languages phonologically similar to the target language tend to be better than results achieved with less similar languages, such that performance degradation is a function of source-target language similarity. Our experiments suggest that a cross-language application of a voice-to-phoneme conversion algorithm is a viable solution to voice-tag recognition for resource-poor languages and dialects. We believe this has important consequences given the globalization of mobile devices and the subsequent demand to provide voice technology in new markets.

5. References

[1] Y.M. Cheng, C.X. Ma and L. Melnar, "Voice-to-phoneme Conversion Algorithms for Speaker-independent Voice-tag Applications in embedded Platforms," *Proc*. IEEE *Automatic Speech Recognition and Understanding* Workshop, Cancun Mexico, 2005.

[2] C. Liu. and L. Melnar "An Automated Linguistic Knowledge-Based Cross-Language Transfer Method for Building Acoustic Models for a Language without Native Training Data" in Proc. *InterSpeech'05*, Lisbon, Portugal, pp. 1365-1368, 2005.

[3] J.J. Sooful, and E.C. Botha, "Comparison of acoustic distance measures for automatic cross-language phoneme mapping," *ICSLP'02*, pp. 521-524, 2002.

[4] Schultz, T. and Waibel, A., "Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme Sets," *Eurospeech '97*, 1:371-373, 1997.

[5] Schultz, T. and Waibel, A., "Polyphone Decision Tree Specialization for Language Adaptation", *ICASSP*. Istanbul, 2000.

[6] T. Holter and T. Svendsen, "Maximum likelihood modeling of pronunciation variation," *Speech Communication*, 29: 177-191, 1999.

[7] F.K. Soong and E.F. Huang "A tree-trellis based fast search for finding the N best sentences hypotheses in continuous speech recognition," in *Proc. International Conf. on Acoustics, Speech and Signal Processing*, pp. 705-708, 1991.

[8] Melnar, L. and Talley, J., "Phone Merger Specification for Multilingual ASR: The Motorola Polyphone Network," *ICPhS 03*, pp. 1337-1340.

[9] Trask, R., *A Dictionary of Phonetics and Phonology*, London: Routledge, 1996, p. 90.