Phone Merger Specification for Multilingual ASR: The Motorola Polyphone Network

Lynette Melnar and **Jim Talley**
Motorola Human Interface Labs, Voice Dialog Systems Lab, USA
E-mail: Lynette.Melnar@motorola.com, Jim.Talley@motorola.com

ABSTRACT

This paper describes the Motorola Polyphone Network (MotPoly), a hierarchical, universal phone correspondence network that defines allowable phone mergers for shared acoustic modeling in multilingual and multi-dialect automatic speech recognition (ML-ASR). MotPoly's organization is defined by phonetic similarity and other language-independent phonological factors. Unlike other approaches to shared acoustic modeling, MotPoly can be effectively used in systems where computational resources are limited, such as portable devices. Furthermore, it is less constrained by language data availability than other approaches. With MotPoly as part of an overall strategy, Motorola's Voice Dialog Systems Lab's ML-ASR team was able to define a set of multilingual acoustic models whose size was only 23% of the largest monolingual model set but whose overall performance was higher than the monolingual models by 1.4 percentage points.

1. INTRODUCTION

ML-ASR technology has faced two principle bottlenecks to successful deployment on small devices. The first potential impasse (independent of platform size) is the cost and availability of language resources. Appropriate language resources are very limited for many languages spoken in small and emerging markets, and the costs and difficulties associated with producing or otherwise acquiring language data can be prohibitive. The second potential impasse is the limited computational resources (memory and processing power) available on portable devices. Since minimization of device costs and power consumption are always high priorities for portable devices, it can be assumed that the available computational resources will continue to be restrictive for some time.

An obvious solution to the language resource and computation costs noted above is to share acoustic models across languages. Sharing phone models helps compensate for deficiencies in language resources and decreases the required overall runtime computational requirements.

2. MOTPOLY: OVERVIEW

MotPoly is a knowledge-based, hierarchically arranged phone merger specification network for shared multilingual and multi-dialect acoustic modeling. The organization of MotPoly is language-independent and phone merger is defined by an internally ranked system of relative phonetic similarity, phonological contrastiveness, and phone frequency. MotPoly functions in ML-ASR as a phone merger framework that constrains data-driven acoustic modeling strategies. Because MotPoly is not biased toward any particular language, language family, or language type, it is a universal, static definition of phone merger and can be used unmodified to specify likely mergers for an unlimited number of phones from any collection of languages or dialects.

MotPoly is of particular value in applications that are restricted in computational and language resources. Because MotPoly specifies mergers of phonetically distinct, but phonologically non-contrastive phones, it limits the inventory size of the multilingual model set beyond that possible with identical cross-language phone merger and it is less reliant on individual language resources through leveraging the combined language assets.

The following subsections discuss the main features of MotPoly design. Subsection 2.1 introduces the Motorola ASCII International Phonetic Alphabet (MAIPA), a feature-based phone labeling system implemented in MotPoly. The structure of MotPoly is identified in subsection 2.2 and the integration of language-independent factors, including phone frequency and phonological universals and tendencies are explicitly discussed. Sections 3 and 4 review MotPoly's application in ML-ASR and provides experiment results. In section 5, concluding remarks are offered.

2.1 MAIPA

MAIPA is a feature-based phone labeling system whose inventory maps to the International Phonetic Alphabet (IPA). [1] Unlike the IPA and other cross-linguistic transcription systems, MAIPA is restricted in form to standard ASCII lower-case alphabetic characters, numerals, and the non-alphanumeric symbol '_' and each

[1] A detailed description of MAIPA is found in [1].

ISBN 1-876346-48-5

label adheres to a strict position class syntax. This design ensures the following three MAIPA features: 1) easily interpretable structure, 2) symbol-consistent extensibility, and 3) straightforward machine processibility (where programming collisions are substantially avoided).

Each individual MAIPA symbol is associated with a specific phonetic feature or feature constellation which depends on the type of the phone represented (consonant or vowel) and its syntactic position in the phone symbol string. The basic character length of any phone representation is two; this obligatory symbol pair is referred to as a phone's *base sign*. In MAIPA, a phone's phonetic features are directly encoded into the label. The first symbol of a base sign unambiguously marks the phone as either a consonant or vowel (i.e. there is no overlap between the consonant and vowel first position class symbols). A consonant label's first symbol also represents the phone's primary place of articulation (POA); its second symbol refers to the phone's primary manner of articulation (MOA) and voicing distinction. A vowel's first symbol represents the phone's backness and roundness in addition to its membership in the vowel class of phones. A vowel's second symbol encodes openness. All non-tonal diacritic symbols (encoding secondary features and, for vowels, stress) are sorted alphabetically behind the base sign and are framed by the marker '_'. Tonal diacritics are suffixed to the right diacritic marker of vowel phone strings.

An example of a simple consonant phone representation, one with only the two obligatory base-sign positions filled, is 'kp', where first position 'k' is associated with the consonant class of phones and indexes 'velar articulation'; consonant second position 'p' indexes 'voiceless plosive'. Thus, 'kp' is a voiceless velar plosive.

The use of MAIPA makes arrangement of phonetic similarity networks more transparent. For example, it is trivial in MAIPA to specify all phones that belong to the class of velar consonants – because all velar consonants, regardless of manner and secondary articulation modifications, are indexed by label-initial 'k'. Thus, 'kp' is a voiceless velar plosive; 'kb' is a voiced velar plosive; 'kn' is a velar nasal; 'kw' is a velar approximate, and so on. Thus MAIPA transparently provides a labeling system consistent with the default phonetic similarity grouping in an encoding like MotPoly, and along with other language-independent factors, determines the correspondence among the phone categories. MotPoly structure is explicitly discussed below.

2.2 MOTPOLY DESIGN

The MotPoly network is a binary branching tree that has as its leaves a nearly exhaustive inventory of MAIPA transcribed phones, derived from a rich set of typologically diverse languages. Nodes on the tree represent phone categories and merger is based upon language-independent (universal) phonological laws and tendencies including relative phonetic similarity, cross-linguistic phone frequency and phonological contrastiveness (see, for example, [2]-[4]). Consider Figure 1 below:

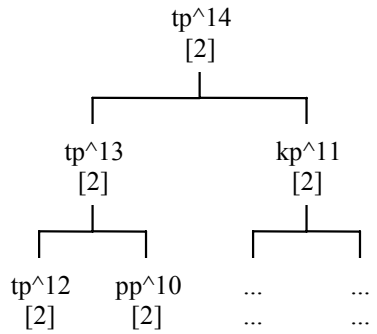

Figure 1: Partial view of MotPoly network

Figure 1 provides a partial view of the MotPoly network. Nodes are characterized by a MAIPA label indexing a phone category that encompasses the phone categories associated with the dependent nodes and leaves. The MAIPA label at each node is followed by a caret character and a number that references the particular sequential reuse of the MAIPA label, beginning with an unmarked '^0' at the leaves. In Figure 1, 'tp^14' is the fourteenth use of 'tp' (IPA 't') and is the parent node of nodes 'tp^13' and 'kp^11' (where 'kp' is IPA 'k').

In comparing leaves and nodes for merger, only the most significant or representative phonetic features associated with the phone categories are considered. These correspond to those features associated with the node's MAIPA label. 'pp^10' (IPA 'p') is a phone category in the network that encompasses not only voiceless bilabial stops, but also other manners and places of labial articulation. Because 'pp' is the most representative phone of the category (and because it would be nearly impossible and otherwise undesirable to compare all phonetic features subsumed in the category) the features used in comparison for merger with 'tp^12 are 'bilabial', 'plosive' and 'voiceless'.

For each node in the tree, a score between '1' and '10' is provided in square brackets, where '10' indicates very high confidence in the merge between the children nodes and '1' indicates very low confidence. Merge scores are derived from three language-independent knowledge sources (KSs), provided below:

KSs of merge scores

1. relative phonetic similarity
2. phonological contrastiveness
3. cross-linguistic inventory frequency

Each KS consists of a local hierarchy of definitions that characterize phone category comparisons. The highest-ranking definition corresponds to the highest merge probability and the lowest ranking definition

ISBN 1-876346-48-5

corresponds to the lowest merge probability for that KS. Phone categories associated with children nodes or leaves are compared and assigned a raw merge score by each of the KSs, depending on how their comparisons match the local definitions. These raw scores are then normalized and this number becomes the actual merge score.

Low merge scores indicate an undesirable merger and characterize the blending of large phone categories in the network. Large phone categories consist of a relatively high number of dependent nodes and leaves and therefore subsume a broad phonetic space. In contrast, high merge scores are associated with the merger of small phone categories and indicate that the merger is desirable – *i.e.*, that important phonetic and phonological distinctions will not be compromised through merger.

Both those factors associated with the KSs and other relevant language-independent factors contribute to the local groupings and hierarchical arrangement of leaves and nodes. The following well-known language universals and tendencies are examples of factors that impact the local MotPoly structure shown in Figure 1:

1. all languages have stop consonants
2. languages tend to have a plain voiceless stop consonant series
3. p, t and k are the most frequently occurring stop consonants in the world's inventories, with t being the most common and p the least common of the three
4. t ⇐ k ⇐ p
5. language phoneme inventories exploit the maxim of sufficient discriminability

The first item is a language universal and points to the relative importance of stops as a consonant class. Other consonant classes (affricates, fricatives, liquids, etc.) are not universally present in the world's sound inventories. Thus among the consonants, stops are ranked higher than the other classes in relative importance. This translates to the idea that the stop class is the highest level consonant class in a universal phone hierarchy. The second item is a language tendency; it provides additional information about stops: in addition to whatever other stop series a language may have - voiced, ejective, implosive, etc. - it is likely to have a plain voiceless series. This means that within the stop class, the plain voiceless series is ranked highest in relative importance. The frequency observation provided in (3) records the overall importance of the three plain voiceless stops 'p', 't', and 'k'. These three stops occur more frequently than any other stops in the world's inventories. And within this set, there is a local frequency hierarchy, where 't' is ranked the highest. The fourth factor, t⇐k⇐p, signifies that if a language has the stop 'p', it must also have the stops 'k' and 't'; and if a language has 'k', it necessarily has 't'. Finally, (5) tells us that contrastive sounds in a given language tend to be distributed proportionally among a range of articulators (for consonants, lips, alveolar ridge, palate, velum, larynx, etc.) and not cluster together at the same region. So, for example, no language contrasts palato-alveolar and palatal phonemes within a given manner series, and contrasts among dentals and alveolars within the same manner series are rare. Notice that the three most common plain voiceless stops, 'p', 't', and 'k' observe this maxim. 'p' is a bilabial stop, 't' is a dental or alveolar stop, and 'k' is a velar stop.

In the partial network view provided in figure 1, 'pp^10' first merges with 'tp^12' to form the phone category 'tp^13'. Since voiceless bilabials are least important typologically among the three stops, 'pp^10' is the first to merge. The second merger provided in this view involves 'tp^13' and 'kp^11'. Because of the two categories, 'kp^11' is less important typologically, it merges with the apical category to form 'tp^14'. Because both of these mergers involves large categories and hence blend important phonetic and phonological distinctions, these mergers are associated with low merge scores.

3 MOTPOLY IMPLEMENTATION

Because the MotPoly network is language-independent in organization, its implementation in acoustic model selection processes should provide the best results for those languages whose phone categories are most typical from a universal perspective. Conversely, those languages that include atypical or less frequently encountered phone categories are hypothesized to be less favorably impacted by MotPoly's use.

MotPoly was implemented by the ML-ASR's Acoustic Modeling team to constrain the data-driven convergence of six languages' model inventories: USA English (EN), Latin American Spanish (ES), Mandarin Chinese (ZH), German (DE), Japanese (JA), and Egyptian Arabic (AR).[2] The training data for these languages were disproportionate, with EN, ES, and ZH having between 50,000 and 70,000 training utterances, while AR, DE, and JA had less than 15,000 utterances. The monophone and triphone model inventory sizes for each of the languages when modeled by itself (*i.e.*, for mono-lingual ASR) are presented in Table 1:

	EN	ES	ZH	DE	JA	AR
Mono	39	35	39	42	32	38
Tri	4210	2309	1866	1019	770	1831

Table 1: Language-specific model inventory

The inventory of the triphone model set was determined based on the training data in each language with the restriction that each model have at least 15 instances for training.

[2] A detailed report of the experiment and results is found in [5].

ISBN 1-876346-48-5

All the data were pooled together across languages and used to train a single set of phonetic context-independent models and a single set of phonetic context-dependent models (see below). The data were processed with a fairly standard MFCC front-end yielding feature vectors with 39 elements (12 cep + E, 13 Δ, and 13 ΔΔ coefficients).

A decision tree technique [6,7] that incorporated MotPoly's phone merger constraints was used to group models across languages. For the multilingual context-independent models, the decision tree was trained based on some language and task ID questions while the context information (traditionally used for decision tree induction) was ignored. This method yielded a context-independent model inventory of 173. To derive the multilingual context-dependent models, context information was included in the decision tree training process. The resulting inventory included 955 models – which is significantly smaller than most monolingual triphone model sets (consider Table 1).

A two pass strategy was utilized in recognition. The first pass employed the context-independent inventory, while for the second pass, the context-dependent inventory was used.

4 RESULTS

Our multilingual models are compared against the language-specific triphone model baselines, and the results are encouraging. On average, the multilingual models outperformed the monolingual sets by 1.4 percentage points – with an inventory size only 23% that of the largest monolingual models. Table 2 summarizes the results, where the white row contains the word accuracy percentages corresponding to the monolingual triphone model sets and the gray row shows the word accuracy percentages corresponding to the multilingual models.

EN	ES	ZH	DE	JA	AR	pooled
68.24	66.72	67.05	53.77	36.23	63.67	64.99
64.75	78.01	67.29	59.10	26.24	61.36	66.39

Table 2: Recognition results

Unsurprisingly, the least favorable results emerged for those languages for which we had restricted training resources and which included marginal, or typologically restricted, phone categories. Thus, AR and JA each have an important atypical consonant series in their respective phoneme inventories. The AR phoneme inventory contrasts non-pharyngealized and pharyngealized consonants while JA contrasts non-palatalized and palatalized consonants. Significantly, none of the other four languages in this experiment include palatalized and pharyngealized consonant series.

In contrast to the above scenario, recognition does improve for those languages associated with typologically more common phoneme inventories, notwithstanding data limitations. Even though our training data for DE consisted of less than 10,000 utterances, significant improvement was attained with the multilingual models due to the convergence of phone categories between DE and the other languages.

The best multilingual results were obtained for ES. We had over 65,000 ES utterances on which to train and the sound inventory of ES includes no marginal phone categories.

5 CONCLUSIONS

The use of a knowledge-based, universal phone merger specification tool such as MotPoly in multilingual model inventory selection can significantly reduce multilingual model inventory size while improving over-all recognition performance. Not unexpectedly, recognition performance improves for those languages principally associated with universally common phone categories while performance shows a degradation for those languages that include a relatively large set of atypical phone categories. For this latter group of languages, it is essential that adequate language resources be available to achieve acceptable recognition results.

REFERENCES

[1] L. Melnar, "The Motorola ASCII International Phonetic Alphabet (MAIPA) and its Extension to the Motorola Polyphone Network (MotPoly)", Motorola HIL technical report, 2002.

[2] R. Jakobson, *Child Language, Aphasia and Phonological Universals*, The Hague & Paris: Mouton, 1941/68.

[3] J. H. Greenberg, C. Ferguson, and E. Moravcsik, eds., *Universals of Human Language, 2: Phonology*, Stanford: Stanford University Press, 1978.

[4] I. Maddieson, *Patterns of Sounds*, Cambridge Studies in Speech Science and Communication, Cambridge: Cambridge University Press, 1984.

[5] C. Liu, "The Test Result for the Multilingual Acoustical Models", Motorola HIL technical report, 2002.

[6] Y. Cheng, "MLite++ Handbook", Motorola HIL technical report, 2002.

[7] Y. Wei, C. Liu, Y. Cheng, and C. Ma, "Language and Task Dependent Allophone Selection for Multilingual Acoustic Modeling and Decoding of ASR Systems", Motorola HIL technical report, 2002.