UNIT SELECTION WITHOUT A PHONEME SET

Alan W Black and Ariadna Font Llitjós

Language Technologies Institute
Carnegie Mellon University
{awb,aria}@cs.cmu.edu

ABSTRACT

With most human languages having less than 1 million speakers it is unlikely that standard commercial systems will be able to justify supporting the vast majority of so-called "minority" languages. In our continuing task of providing tools for building synthetic voices in currently unsupported languages, this paper describes a number of experiments in building synthetic voices without requiring specific phonetic knowledge of the target languages. Even when a language is well studied defining an appropriate phoneme set is never easy. The work presented here shows the adequacy of unit selection synthesis techniques when no explicit phoneme set is available.

1. BACKGROUND

In the continuing goal to provide sufficient tools to build synthetic voices for all languages, this paper describes experiments in restricting phonetic knowledge in building voices. As the technology improves, we are finding more and more uses of speech output that were not considered before. There are around 6,000 active languages in the world and it seems unfair to exclude them from spoken language output systems because they are not one of top 20 or so langauges by population or economics.

In fact, we believe that speech technology can be most helpful when dealing with minority languages. In languages where there is a low level of literacy, either because reading/writing is taught in some other more widely spoken language, or because of the lack of educational resources, a spoken language system may be the only reasonable way to distribute information.

The AVENUES project at CMU is concerned with building speech to speech translation systems for indigenous languages in South America. This project is designed to address issues in building speech and translation components even when very little data exists. It is not unusual in minority languages that the orthography is not well defined.

[1] gives a description of writing systems used through out the world and their relative opacity with respect to their phonetics. It is our belief that languages with a short history in writing are often more closely related to their phonetics than those with a longer history. However there may also often be the complication that the alphabet used for such minority languages is not appropriate, as it may not have the variation suitable for the language. For example, the Spanish alphabet may be used for a native American language, and the shortcomings may be resolved by the addition of diacritics. That is the case of Mapudungun, an indigenous language spoken by around one million people in Chile, which uses umlaut in addition to an alphabet based on standard Spanish. Importantly, even with a defined alphabet, the relationship between the orthography and the phonetics may not actually be one to one. Especially when one considers dialects.

Making the assumption that there will be a relationship between the letters and pronunciation, we have built a number of synthesizers which use letter information alone to determine the "phone" set.

To build these voices, we based our techniques on the framework provided by the CMU FestVox tool suite [2], which provides basic templates and tools for building synthetic voices in new languages. It has already been used to build a wide range of voices in at least 40 different languages.

2. AN EXPERIMENT

Our basic experiment involved taking recordings of two different dialects of Spanish. Spanish was chosen as the language for testing, even though it is a well defined case, as we have already built Spanish synthesizers before. It is that familiarity that made us choose it for this experiment.

Also the relationship between the writing system and phonology is relatively close. However, although it is not complex, it is is also not simply a one-to-one relationship.

In order to build a voice without using phonetic information we used the letter set as the phoneme set. Thus our phoneset consists of 26 standard English letters plus the accented characters á, é, í, ó, ú, and ñ, and also SIL (silence). We did use our knowledge of the language and made all letters lower case, and omitted rarer accented characters like ç.

The texts we used for recording had been automatically

selected from various newspaper texts to give best diphone coverage, for a general Spanish synthesizer. More elaborate selection techniques, such as [3] were not available to us as they would require a more detailed phonetic and acoustic analysis of the language. However we are aware that our data used in our recordings did use some phonetic knowledge in its construction, but still feel the basic experiment is valid.

The lexicon, the process that provides pronunciations from words, simply takes each word, converts the characters in it to lower case and returns them as a list of phones. As no vowel/consonant information is available each word is coded as a single syllable.

Another knowledge-based expansion of the data is conversion of numeric strings to number words, as is conventional in all text to speech synthesizers. As our text was selected from newspapers, a number of digit strings and abbreviations appeared in the text. Such tokens do not have a closely related pronunciation to their letter sequence. In a standard Spanish synthesizer token expansion rules are used to expand these "non-standard words" to explicit, complete words. For this experiment, we used the same expansion set for the data, thus using some knowledge of the language. However, this is equivalent to requiring each word to be written in full.

The prompt list of 419 utterances was recorded by a female Castillian Spanish speaker and by a male Colombian Spanish speaker. The number of words is 5044, and the number of units in these databases is around 28,000. The exact number of units varies between speakers due to the number of leading, trailing and inter-phrasal SIL phones required as the speaker did not deliver the data at exactly the same speed, nor with the same phrasing. The data was recorded in professional recording studios, at 16KHz samples with a simultaneous EGG (larygnograph) channel.

3. LABELING THE DATA

In previous synthesizers we have labeled spoken prompts by using DTW (dynamic time warping) techniques on a synthesized version of the prompts generated by an existing synthesizer. This technique, based on [4], works well within a language but we have also often used this cross-lingually. In the latter case, one needs to take a close language (or perhaps just English) and map the phones in the new language to approximations in the target language. Synthesizing using that mapping provides acoustic prompts, which although may sound very English, have approximately the right properties to allow reasonable alignment using DTW.

However, such techniques require phonetic knowledge to decide which phoneme in the labeling language maps to which in the target language. And we wish to require no such knowledge of the target language.

In this case, we used the SphinxTrain acoustic modeling tools [5] to build context-dependent semi-continuous HMM models using the letters as phone names. This does require an orthographic transcription of the prompts (which were read by the native speaker when they were recorded). It also implicitly requires sufficient data to have reasonable acoustic coverage.

At this point, we have probably taken advantage of some phonetic knowledge in the original choice of sentences to include in the prompt set, in that they were selected to have a rich diphone coverage. However it could be argued that using a selection criteria based on letter rather than phone distribution would produce a similar database.

4. CLUSTER BASED UNIT SELECTION SYNTHESIS

The unit selection technique is that described in [6]. In this technique, units of the same type are collected together and an acoustic distance is calculated between each occurrence. A recursive splitting algorithm is used to find which high level questions can be used to split the data such that the mean acoustic distance between members of the partition is minimized. Thus clusters of acoustically similar units are indexed by trees of high level questions.

More formally, we define the acoustic distance $D(U,V)$ between two units U, and V where $|V| > |U|$ as

$$P\frac{|U|}{|V|}\sum_{i=1}^{|U|}\sum_{j=1}^{n}\frac{W_j}{n\sigma_j|U|}abs(F_{ij}(U) - F_{(i\frac{|V|}{|U|})j}(V))$$

where P is a duration penalty, $|U|$ is the number of frames in U, W_j is the weight for parameter j. $F_{xy}(U)$ is the parameter y of frame x of unit U, σ_j is the standard deviation of parameter j, and there are n parameters. The term $F_{(i\frac{|V|}{|U|})j}$ is F_{xy}, where the x index is computed as $i \times \frac{|V|}{|U|}$, and $y = j$.

We can then define the impurity of a cluster as

$$Impurity(C) = \frac{1}{|C|^2}\sum_{i=1}^{|C|}\sum_{j=1}^{|C|}D(C_i, C_j)$$

Then, using standard CART techniques, we greedily find the question that gives the best information gain, and split clusters to minimize the summed impurity of the sub-clusters.

The acoustic distance between each unit is calculated from the mahalanois euclidean distance between pitch synchronous vectors of Mel cepstrum coefficients plus coefficients for duration and F0.

This method is designed to automatically distinguish between acoustically distinct units based on context. It is this particular factor that we are exploiting in this case. As we are assuming no phonetic knowledge, the acoustics and letter contexts (plus higher level information) are being used to define the units that will be selected at run time.

5. EVALUATION

Three levels of tests were carried out. The first was within a particular dialect to confirm that pronunciations of letters in different context were properly treated.

Looking at the decision cluster trees, we can see that letter context (and position in word) is being used to differentiate between the multiple realizations of a letter-phone. For example, both voices managed to learn the distinction between the 3 different ways to pronounce "c" (/k/, /ch/ and /th/ or /s/, depending on the dialect) and the 2 different ways to pronounce "g" (/g/ and /j/).

Word	Castillian	gloss
casa	/**k** a s a/	house
cesa	/**th** e s a/	stop
cine	/**th** i n e/	cinema
cosa	/**k** o s a/	thing
cuna	/**k** u n a/	cradle
hechizo	/e **ch** i th o/	charm, spell

In Spanish the letter "c" may be pronounced /k/, /ch/ and /th/ or /s/ (depending on dialect). The choice of phone is determined by the letter context.

Other examples of how these letter-based voices learned context sensitive differences can be found in the sequences "que-", "qui-", "gue-" and "gui-", where the "u" does not get pronounced (*quien, querida, guerra, guitarra*), and in the single "r" when it apears at the beginning of the sentece, pronounced /rr/ (*rosado*) as opposed as when it appears in an intervocalic position, pronounced /r/ (*coral*).

This shows that given enough audio examples of all these different contexts, the synthesizers were able to learn context-sensitive differences, and thus not knowing what the phoneme set of a language is, it is still possible to build a voice for that language.

The letter "x" performed worst as the systems seem to always pronounce it as /ks/ but in many cases it should be pronounced as /s/. This may be caused by the relatively rare occurrence of the letter in the database, only 52.

The second level of evaluation we investigated is how dialect differences are reflected. The most obvious difference between Castillian Spanish and Colombian Spanish is the use of /th/ and /s/ for the letters "c" and "z".

Word	Castillian	Colombian	gloss
caza	/k a **th** a/	/k a **s** a/	hunting
cesa	/**th** e s a/	/**s** e s a/	stop
cine	/**th** i n e/	/**s** i n e/	cinema
hechizo	/e ch i **th** o/	/e ch i **s** o/	charm, spell

Dialectal differences for the letters "c" and "z" captured correctly by our two voices

The third evaluation was less specific to particular identifiable phenomena, and focused on the overall synthesis quality. Two short paragraphs were taken from *La Vanguardia* (May 20, 2002) and were synthesized by each of the two voices.

> Sevilla, Agencias. Los sindicatos UGT y CC.OO. han exigido al presidente del Gobierno, José María Aznar, que convoque la mesa de negociación de la reforma del sistema de protección por desempleo, tras reunirse con el presidente de la Junta de Andalucía, Manuel Chaves, y el de Extremadura, Juan Carlos Rodríguez Ibarra.
>
> El secretario general de UGT, Cándido Méndez, junto al responsable de CC.OO., José María Fidalgo, reiteró la necesidad de que sea Aznar quien convoque y esté presente en esta mesa, si bien precisó que esta reunión no servirá para nada si la cita no comienza con el anuncio del Gobierno de que retirará su actual propuesta de reforma.

This passage consists of 109 words. The synthesized versions from the Castillian and Colombian letter based synthesizers were listened to by a native Spanish speaker (who is the Castillian speaker and an author of this paper). Each word was assigned a value of good, poor or bad.

Dialect	good	poor	bad	% good
Castillian	102	6	1	93.57%
Colombian	99	5	5	90.82%

Where "poor" is defined to be words that are not clearly synthesized. An example of token that was labeled as "poor" is *CC.00.* which was not analyzed properly and thus it was assigned a default error word pronunciation.

It should be noted that as no hand correction to the labels were done and some of these errors are due to more conventional unit selection errors than to the letter/phone restrictions that we are imposing on these particular builds. Hand correction of segmental boundaries is always worthwhile in a unit selection synthesizer but at this stage we did wish to introduce that complication in this experiment.

The one phonetic error in the Castillian voice was "Sevilla" pronounced as /s e v i l a/ rather than /s e v i y a/.

The Colombian voice also made the same error in "Sevilla" and actually pronounced as /s/ an instance of a "c" which should be pronounced as /k/ (actual – /asetual/, where the "e" is probably due to bad alignment). The other bad examples may be better attributed to bad alignments (as were all extra inserted vowels).

6. DISCUSSION

Fully automatic builds of synthesizers in unresearched languages is a long way off, however with the greater demand for support in minority languages it is something that should be addressed.

Using acoustic information to find distinctions is implicitly what we have been trying to do in unit selection synthesis, thus explicitly taking advantage of that should not be a surprise.

Anecdotal evidence of this already shows up in other synthesizers build by us. When using an American English based synthesizer with US English phoneset, a US English lexicon, and a Scottish English speaker, the lexical entries do not properly match the speaker's pronunciations. For examples palatalized /uw/ as found in British English in /t y uw z d ey/ (Tuesday) is defined as /t uw z d ey/ in the US English lexicon. When this labeling is used against a Scottish English speaker the /y-uw/ segment is labeled as /uw/. Thus when other words are synthesized with similar contexts the palatalization is still generated thus words labeled as /s t uw d eh n t/ (student) may correctly, for the dialect, be synthesized as acoustics that could be labeled as /s t y uw d eh n t/.

It should be noted that it is rare that absolutely no phonetic knowledge is available for a language and often at least some information (vowel/consonant) can be directly derived from the orthographic system. However it is not unusual that there are no linguistically knowledgeable speakers of the language available, and native speakers are often not explicitly conscious of the distinction they are making. In a practical sense, a gross classification of phonemes can be reasonably specified but fine distinctions are much harder.

It is worth comparing the complexity of mapping letters directly to acoustics, with the more standard approach of having an intermediate finite phone set. As we are considering mapping without explicit lexicons it is best to compare with the automatic letter to sound rule mappings as described in [7]; in this case, we map letters to predefined finite phone sets. Importantly, letter to sound training sets are bigger, because it is easier to collect text than speech. However the difference in size is only perhaps one order of magnitude (5000 words vs. 50,000 words), and in the letter to acoustic case we have selected data deliberately to get coverage.

Machine learning techniques could allow us to assume a hidden layer that explicitly represents a phone set, but we have not investigated that yet.

Another direction that may be worth investigating is to cluster the acoustics independent of any labeling and then match the types identified by the clusters to letters. Such techniques for acoustically derived units have been studied for speech recognition (e.g. [8]) but have not yet been investigated for unit selection synthesis.

It is clear that depending on the language and knowledge available, there is a scale of pure letter to acoustic through to letter to phone and phone to acoustic models. But we would like to make that scale available to the voice builder so they may best take advantage of the information they currently have available.

Another point that we wish to make clear is that without native speaker's feedback for evaluation the ultimate quality of a synthetic voice cannot be determined. As those who work in the field immediately notice, synthesis in languages you are not familiar with typically sound better than synthesis in languages you are knowledgeable about. It requires fluent speakers to properly evaluate content. In our experience in building synthesizers for minority language we find, anecdotally, that listeners can be more extreme that those in more common languages. On one hand, that there is a synthesizer at all in their language can make some native listeners accept what is not the best possible synthesis. On the other hand, listeners of minority languages are likely to be unfamiliar with speech synthesis, and they can even find listening to high quality recorded speech difficult to understand.

7. ACKNOWLEDGMENTS

We are grateful to Cepstral, LLC for providing recording studio time and support for collecting the databases used in this work. We also like to thank German Torres for helping us record the Colombian Spanish data. This work was funded in part by NSF grant 0121631 Avenues. The opinions expressed in this report do not necessarily reflect those of NSF.

8. REFERENCES

[1] R. Sproat, *A Computational Theory of Writing Systems*, Cambridge Univiversity Press, 2000.

[2] A. Black and K. Lenzo, "Building voices in the Festival speech synthesis system," http://festvox.org, 2000.

[3] A. Black and K. Lenzo, "Optimal data selection for unit selection synthesis," in *4rd ESCA Workshop on Speech Synthesis*, Scotland., 2001.

[4] F. Malfrere and T. Dutoit, "High quality speech synthesis for phonetic speech segmentation," in *Eurospeech97*, Rhodes, Greece, 1997, pp. 2631–2634.

[5] Carnegie Mellon University, "SphinxTrain: building acoustic models for CMU Sphinx," http://www.speech.cs.cmu.edu/SphinxTrain/, 2001.

[6] A. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Eurospeech97*, Rhodes, Greece, 1997, vol. 2, pp. 601–604.

[7] A. Black, K. Lenzo, and V. Pagel, "Issues in building general letter to sound rules," in *Proc. ESCA Workshop on Speech Synthesis*, Australia., 1998, pp. 77–80.

[8] M. Bacchiani, M. Ostendorf, Y. Sagisaka, and K. Paliwal, "Design of a speech recognition system based on acoustically derived segmental units," in *ICASSP-96*, Atlanta, Georgia, 1996, vol. 1, pp. 443–446.