Speaker Clustering for Multilingual Synthesis

Alan W Black and Tanja Schultz

Language Technologies Institute
Carnegie Mellon University, Pittsburgh, PA, USA
awb@cs.cmu.edu and tanja@cs.cmu.edu

Abstract

Today, speech synthesizers in new languages are typically built by collecting several hours of well recorded speech in the target language. The time and effort involved in collection and correction can be prohibitive when lack of resources is common in addressing under-represented languages. An alternative method is to use acoustic data from an existing synthesizer in a different language and train adaptation models from a small corpus (20-50 sentences) in the target language. Following the work of GlobalPhone [Schultz, 2001], which uses multi-lingual databases and adapts acoustic models to target languages for speech recognition. This paper presents one aspect towards a solution of building monolingual synthesizer when little or no target language data is available. Our particular focus is in selecting appropriate speakers from large number of candidates through voice clustering techniques.

1. Synthesis and Recognition

1.1. Multilingual Phone Sets for Speech Recognition

Our research in the design and implementation of a global phoneme set for speech recognition and speech synthesis is based on the assumption that the articulatory representations of phonemes are so similar across languages, that phonemes can be considered as units which are *independent from the underlying* language. Based on this assumption the language specific phoneme inventories of languages can be unified into one global set. This idea has been embodied for example in research on language identification [Andersen et al., 1997], [Corredor-Ardoy, 1997].

In our multilingual speech recognition work [Schultz, 2001] we defined a global unit set for 12 languages (Chinese, English, French, German, Japanese, Korean, Croatian, Portuguese, Russian, Spanish, Swedish, and Turkish) based on the IPA scheme [IPA, 1993] and developed acoustic models for speech recognition. Sounds of different languages, which were represented by the same IPA symbol, shared one common unit in this global acoustic model set. According to this idea we differentiate between the group of language independent *poly-phonemes* containing phonemes occurring in more than one language, and remaining language dependent *mono-phonemes*. Table 1 summarizes the poly-phonemes and mono-phonemes for 12 languages.

For each poly-phoneme the upper half of Table 1 reports the number of languages which share one phoneme. The lower half of Table 1 contains the number and type of mono-phonemes for each language. In total, the global unit set consists of 485 language dependent phonemes which had been shared into 162 classes. Therefore, on average, each phoneme of our global unit set is shared by 3 languages. The phoneme *share factor* increases with the number of languages and strongly depends on the involved languages, implying that the phoneme inventories of some language subsets are quite similar while others are not. The global unit set in conjunction with the acoustic models covering 12 languages of the world provides us with the optimal basis to select phonemes for new languages and use the corresponding language independent acoustic models as seeds for speech recognition and speech synthesis when targeting new languages.

Shared by	#	Modeled Phonemes (IPA symbols)	
	83	Polyphonemes shared across ≥ 2 languages	
		Consonants	Vowels
All	4	m,n,s,l	-
11	7	p,b,t,d,k,g,f	-
10	3	-	i,u,e
9	6	ŋ,v,z,j	a,o
8	1	ʃ	-
7	3	r,h,tʃ	-
6	1	-	ɛ
5	9	ɲ,ʒ,x,ts,dʒ	i:,y,ə,ɔ
4	4	-	ɨ,ø,ɑ,ei
3	11	ɾ,w,ç	ɪ,u:,e:,œ,o:,æ,ai,aʊ
2	34	p^h,t^h,d^j,k^h,g^j,ʁ,$ɾ_r$, θ,ð,s^j,z^j,ʂ,ɻ,ts^h,$tʃ^j$	ˈi,y:,ɯ,ʊ,ˈe,ɛ:,ø:,a:,ˈa,ɑ:, ˈu,ˈo,aɪ,au,ia,io,eu,oi,oʊ
	79	Monophonemes belonging to *one* language	
		Consonants	Vowels
CH	15	tʂ,$t^hʂ$,cç,$cç^h$	iʊ,iɛ,ua,uɛ,uɔ,ya,yɛ, iao,uɛi,uai,ioʊ
EN	5	$ɾ_d$	ʌ,ɝ,ɔi,ə˞
FR	5	ɥ	ɛ̃,œ̃,ɑ̃,ɔ̃
GE	3	-	ɐ,ʏ,ɔʏ
JA	2	ʔ	ɯ:
KO	14	$p^ʔ$,p',$t^ʔ$,t',$k^ʔ$,k', s',$c^{ʻh}$	ie,iɘ,iu,ɨi,oa,uɘ
KR	1	$dʒ^j$	-
PO	8	-	ĩ,ũ,ɛ̃,õ,ɐ̃,ew,ow,aw
RU	15	p^j,b^j,t^j,m^j,r^j,v^j, f^j,$ʒ^j$,l^j,$ʃ^j$,$ʃ^j$,$tʃ^j$	ja,jɛ,jɔ,ju
SP	2	β,ɣ	-
SW	9	ʈ,ɖ,ɳ,ɭ,ks	œ:,æ:,ʉ:,ɵ
TU	0	-	-
∑	162	Silence and noises shared across languages	

1.2. Multilingual Synthesis

With the demand for speech synthesis coverage of new languages rapidly increases we must continue to improve

techniques to deliver high quality understandable synthesis while using as few human resources as possible. A number of systems have been developed to provide well defined tools to build support in new languages (FestVox [Black and Lenzo 2000], IBM Trainable Speech Synthesis System [Donovan et al. 1998]). Those systems however have depended on the collection of a large carefully collected database of at least 500 sentences, but typically up to 10 times that size or more. These databases are then carefully labeled and tuned to produce good results.

However in our experience with people using our FestVox development tools, which contains substantial documentation and well designed scripts, are not always sufficient for non-experts to build support for new languages. Although it is easy to blame the complexity and inadequacy of our tools, it is more productive to investigate why collecting large high quality databases is so hard, and to find more tolerant methods that will lead many more people to success.

One clear observation gained from building many different synthetic voices is that most people do not have consistent enough voices to deliver 5000 sentences well enough to build a good synthesizer. This is of course why commercial voice building involves professional voice talent. We also observe that in building speech recognition acoustic models, multi-speaker and more natural delivery is required in order to model the variability that we find across speakers.

In order to utilize speech recognition data for speech synthesis it is clear that the existing methods of unit selection are unlikely to succeed given the wider variety of data and less consistency within it. With the work on HMM-generation synthesis [Tokuda et al., 2000] we are seeing a move away from high quality instance selection, where appropriate sub-word units of natural speech are selected from large carefully recorded databases, to techniques that combine multiple instances into a parametric model from which the speech is then generated.

[Latorre et al. 2005] have also investigated using multilingual data for monolingual synthesis. In this paper we extend existing work by investigating automatic speaker selection techniques for monolingual and multilingual synthesis and perform this on a larger variety of languages.

2. Unit Selection vs Parametric Synthesis

Since around the 1980s, concatenative synthesis techniques have gradually overtaken the earlier format synthesis techniques. First, diphone synthesis became the accepted method for building high quality voices where phone sized chunks (from the middle of one phone to the middle of the next) were carefully recorded and labeled to provide a well-defined fixed inventory of speech units. The inventories of such systems grew, explicitly adding more variation. With larger inventories it became harder to ensure complete coverage (and have a voice talent properly deliver the desired phonetic realization). Therefore a more general technique of automatic selection of sub-word units from large databases of natural speech has evolved. Unit selection techniques like that of [Hunt and Black 96] and [Donovan and Woodland 95], employed acoustic measures to select appropriate sub-word units from large databases. Unit selection techniques have been very successful at producing very high quality speech, but at the cost of requiring good large databases. The results model the speech style in such databases and when synthesizing outside that style or coverage is required these voices can quickly become sub-optimal. In order to keep their high quality it is typical to do little or no prosodic modification, again limiting synthesis to within the coverage of the databases. Also although these voices are typically very good, there always remains a possibility that some bad units (improperly spoken or labeled) may be selected and hence cause sub-optimal synthesis.

As we look for more controllable synthesis, alternative methods are being investigated. [Tokuda et al., 2000] first displayed a parametric synthesis technique were sub-phonetic segments are modeled not as sets of instances of units (as in unit selection) but parametric models which are used to statistically generate speech. One could crudely view such clusters no longer as sets but as averages of the instances. This method has been show to provide high quality understandable speech [Bennett, 2005] and language independence [Tokuda et al., 2002]. More recently Latorre [Latorre et al., 2005] has investigated how multi-lingual databases such as GlobalPhone may be used within HMM-generation synthesis.

3. CLUSTERGEN Parametric Synthesizer

The CLUSTERGEN synthesizer is a new synthesis technique added to the FestVox suite of voice building tools. Specifically it offers a clustering technique for HMM-state-sized segments. The training data consists of natural speech data labeled with an HMM-based automatic labeling system. Such a system is included with FestVox but such labels may be generated by any other system.

CLUSTERGEN depends on a reversible (analysis/synthesis) parameterization of speech. In this work we use MELCEP analysis and an MLSA filter for resynthesis [Imai 1983] which is the same analysis/synthesis methods used in NITECH's HTS.

Order-24 MELCEP feature vectors from the same HMM-state sized segments are clustered using the *wagon* CART tree builder. The features used for tree building are the articulatory features from the IPA phoneme definition as well as other phonetic, syllabic and contextual features. The clusters are optimized to minimize the sum of the standard deviations of each MELCEP feature multiplied by the number of samples in the cluster. The MELCEP features are not normalized thus as the magnitudes of the MELCEP features vary from C0-C23 this will bias the clustering to minimize lower feature variation.

At synthesis time the desired phones are generated as for other synthesis techniques. Then for each HMM-state feature vectors are created for the duration of each HMM-state, then the appropriate CART tree is used to select an appropriate cluster. The means of each parameter in the vectors are selected as the feature values. Apart from short term

smoothing no other delta information is used in the current system.
F0 information is generated in a set of parallel CART trees in a similar way. This largely follows the basic form of HMM-generation synthesis.

4. Dataset, Labeling, and Clustering

4.1. GlobalPhone Database

For our experiments we used the multilingual database GlobalPhone [Schultz, 2002] that was collected for the purpose of developing and evaluating large vocabulary continuous speech recognition systems in the most widespread languages of the world. GlobalPhone is designed to be uniform across languages with respect to the amount of text and audio per language, the audio data quality (microphone, noise, channel), the collection scenario (task, setup, speaking style etc.), and the transcription conventions. To date, the GlobalPhone corpus covers 18 languages Arabic (Modern Standard Arabic), Bulgarian, Chinese-Mandarin, Chinese-Shanghai, Croatian, Czech, French, German, Japanese, Korean, Polish, Portuguese (Brazilian), Russian, Spanish (Latin American), Swedish, Tamil, Thai, and Turkish. The data acquisition was performed in countries where the language is officially spoken. In each language about 100 adult native speakers were asked to read about 100 sentences each. The read texts were selected from national newspaper articles available from the web to cover a large vocabulary. Speech data was recorded with a Sennheiser 440-6 close-speaking microphone and is available in same characteristics for all languages, i.e. PCM encoding, mono quality, 16bit quantization at 16kHz sampling rate. The entire GlobalPhone corpus contains over 350 hours of speech spoken by more than 1700 native adult speakers.

Since GlobalPhone was originally designed for speech recognition, the coverage of a large number of speakers was prioritized over coverage of a large number of utterances per speaker. The latter would be more desirable for the purpose of speech synthesis. However, as the goal of this research is to perform rapid adaptation on a small set of utterances, GlobalPhone is an appropriate choice. As GlobalPhone has been applied to other multilingual TTS work [Latorre et al. 2005] it could be established for benchmarking.

4.2. Speaker Selection and Clustering

One important aspect when merging models for synthesis across speakers and languages is to define a homogenous set of similar speakers. We investigated procedures for speaker selection, a manual and an automatic one, and compare the results of both procedures with respect to synthesis performance. In the manual selection process an expert hand-selected one speaker per language by listening to a large number of speakers and picking the best matching voices within one gender category. Among the matching speakers, those with the most sentences recorded were selected. The automatic procedure consists of a two-step approach. In the first step we used speaker characteristics that have a strong impact on the voice, i.e. gender, age, and smoking preference that are available for each GlobalPhone speaker. Data analysis showed that the group "Non-smoking females between 18 and 24 years" has the most balanced distribution among languages. In total 213 speakers from 10 languages fall into this group. In the second step we randomly selected 6 utterances (around 50 seconds) per speaker to perform an automatic speaker clustering.

We applied a hierarchical, agglomerative clustering technique described in detail in [Jin and Schultz, 2004]. At the beginning, each utterance is considered as one cluster, resulting in 1278 different clusters (213 speakers x 6 utterances). At each clustering step, we computed the pair-wise distances between all segments and the two segments with the smallest distance were merged. The distance between two segments Win_1 and Win_2 is defined as:

$$D(Win_1, Win_2) = -\log \frac{P(X_c|\theta_C)}{P(X_A|\theta_A)P(X_B|\theta_B)}$$

where θ_A, θ_B, and θ_C are Gaussian Mixture Models (GMMs) for the data in Win_1, Win_2, and the combination of Win_1 and Win_2, respectively. To get more reliable model estimates, we trained one Tied GMM (TGMM) using the data of all given segments and then built segment-specific GMMs by adapting the TGMM using the data from that segment only. As a stopping criterion for the clustering process we applied the Bayesian Information Criterion [Chen and Gopalakrishnan, 1998].
As a result the 1287 utterances spoken by 213 speakers from 10 languages were clustered into 6 clusters. Only one of these clusters contained speakers from all 10 languages. From this cluster we selected one speaker per language by applying the following criteria (1) all 6 utterances of the speaker fall within the same cluster, (2) the speaker belongs to the training set, and (3) among the remainder speaker select the one with the longest speech segment duration.

4.3. Data Labeling

The labels for the manual and cluster-based selected speakers were automatically generated by forced alignment using the Janus Recognition Toolkit (JRTk). The recognizers for all 10 languages had been formerly trained based on the GlobalPhone database. All recognizers use the same pre-processing, HMM topology, and roughly the same acoustic model size. We trained fully continuous 3-state HMMs with 2000 to 3000 quintphone models using 32 Gaussians per state. The 13 mel cepstral coefficients, power, and the first and second derivatives had been reduced to 32 dimensions using Linear Discriminant Analysis. The performance of the recognizers in all 10 languages ranges between 10% and 20% WER [Schultz, 2001]. Since most of the above selected speakers belong to the training sets of the recognizers, the label quality is expected to be reasonably good. Nevertheless, the different performances of the recognizers may have an influence on the quality of the labels. In general it is assumed that the automatically generated labels have a poorer quality than those produced by human expert labelers. The latter procedure is typically used for high quality synthesis.

5. Basic Experiments

Using the basic ten selected speakers we build a CLUSTERGEN synthesis voices for each language (MONO) and a combine voices from all the languages (MULTI).
In each case we tested the constructed models on a set of held out sentences from the contributing speakers. The sentences were not part of the training set. The number of test sentences was approximately 10% of total number of sentences available from that speaker.

The quantitative measure used is Mel Cepstral Distortion [Toda et al 2004] which has been used to for voice conversion, it is a spectral based distance.

$$10/\ln 10\sqrt{2\sum_{d=1}^{24}\left(mc_d^{(t)}-mc_d^{(e)}\right)^2}$$

5.1. Monolingual vs Multilingual TTS

The quality of the generated synthesis varies from language to language. The English and German synthesis is completely understandable, even if it sounds a little buzzy. The other languages too are basically understandable.

LANG	MONO	MULTI	MULTI+
CH	7.60	7.72	7.53
DE	5.56	7.02	6.43
EN	5.41	7.08	6.71
JA	5.55	6.63	6.17
CR	6.73	7.32	6.75
PO	6.87	7.47	7.06
RU	7.15	8.60	7.54
SP	6.30	8.67	7.99
SW	6.04	7.23	6.77
TU	6.92	8.97	8.48

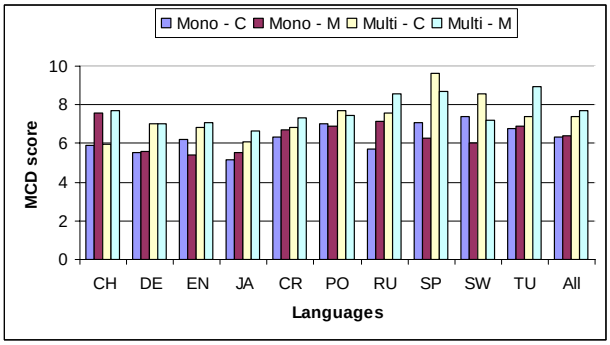
The above table shows results for the manual selected speakers. The languages are: Chinese (CH), German (DE), English (EN), Japanese (JA), Croatian (CR), Portuguese (PO), Russian (RU), Spanish (SP), Swedish (SW) and Turkish (TU). MONO is where each synthesizer consists of only data from the target language. MULTI is where all languages are used for multilingual TTS. MULTI+ includes a feature which identifies the language from which the samples came from.

LANG	MONO	MULTI	MULTI+
CH	5.87	5.94	5.86
DE	5.53	7.03	6.06
EN	6.20	6.85	6.44
JA	5.17	6.10	5.49
CR	6.34	6.85	6.51
PO	7.02	7.72	7.20
RU	5.71	7.55	6.42
SP	7.08	9.63	8.04
SW	7.40	8.55	8.04
TU	6.76	7.41	6.95

This second table shows the results for the speaker selected through clustering, which is discussed more fully in the following section. What can be clearly seen is that the MONO versions of the voices produce the best results. Including multi-lingual data does not improve the results, except for Chinese which actually has the smallest intersection of phones with the other languages. Adding explicit language information (MULTI+) improves the results significantly for most languages compared to MULTI. But it is still not clear how using non-target language data boost results. Ultimately we wish to reduce the amount of required data in the target language. Although we carried out no formal listening tests, the synthesized results are mostly understandable by native speakers, though in all languages the amount of target language data is much smaller than one would usually desire in unit selection synthesis databases.

5.2. Manual Selection versus Speaker Clustering

Figure 1 compares the results of the manual speaker selection (M) to the automatically clustered speaker selection (C) for both, the monolingual MONO and the multilingual TTS MULTI. Overall it shows that the selection by clustering improves the performance slightly in both, the monolingual and the multilingual case (category "MULTI"). As expected, the multilingual TTS benefits more from the clustering than the monolingual TTS does, since in multilingual TTS the homogeneity of the speaker group becomes a more important issue. For most languages we see a moderate (EN, JA, CR) or even significant improvement (CH, RU, TU) by using speaker clustering for multilingual TTS. However, for PO we see a slight, and for SP and SW a significant degradation. One reason could be that in case of Portuguese and Spanish the amount of training data was larger for the manual selected speaker (SP: 12 min vs 7 min, PO: 13.6 min vs 3.4 min), however the performance differences in PO are smaller than in SP even so PO had the larger differences in data. Furthermore, we had similar discrepancies in case of DE (11.7 vs 6.4 min) which did not have any impact on the performance. For SW the amount of data is the same for both speakers. In future work we are planning to control the amount of training data. Another aspect is the quality of labels that are used for synthesis. As mentioned above all synthesis experiments were performed with automatically generated labels. As a consequence the label quality depends to some extend on the speech recognition performance. Among the given recognizers the PO, SP, and SW have the worst performances.



6. Conclusions

It is clear that the variability across languages is difficult for reasons such as phoneme perplexity, amount of data, label accuracy, and appropriateness of speakers, as well as the homogeneity among the speaker set. Our results indicate that for most languages the automatic selection of speakers outperform a manual selection and allows better quality synthesis. In future work we plan to extend the list of speakers to more than one per language and investigate how the quality of synthesis varies with the amount of speakers and target speech.

7. Acknowledgements

This work is in part supported by the US National Science Foundation under grant number 0415021 "SPICE: Speech Processing --- Interactive Creation and Evaluation Toolkit for new Languages." Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

8. References

[Andersen, 1997] Andersen, O. and Dalsgaard, P., *Language Identification based on Cross-language Acoustic Models and Optimised Information Combination*. In: Proc. Eurospeech, Rhodes 1997, pp. 67-70.

[Bennett, 2005] Bennett, C. *Large Scale Evaluation of Corpus-Based Synthesizers: Results and Lessons Learned.* In Proc Interspeech, Lisboa, Portugal, 2005.

[Black and Lenzo 2000] Black, A. and Lenzo, K. *Festvox: Building Synthetic Voices.* http://festvox.org

[Corredor-Ardoy, 1997] Corredor-Ardoy, C., Gauvain, J.L., Adda-Decker, M. and Lamel, L., *Language Identification with Language-independent Acoustic Models*. In: Proc. Eurospeech, pp. 355-358, Rhodes, Greece, 1997.

[Chen and Golapakrishnan, 1998] Chen, S.S. and Golapakrishnan, P.S., "Clustering via the Bayesian Information Criterion with Applications in Speech Recognition", ICASSP 1998.

[Donovan and Woodland, 1995] Donovan, R. and Woodland, P. *Improvements in an HMM-based speech synthesizer*. In Proc. Eurospeech pp 573-576, Madrid, Spain, 2005.

[Donovan and Eide, 1998] Donovan R, and Eide E. *The IBM Trainable Speech Synthesis System.* In Proc 1CSLP, Sydney, Australia, 1998.

[Hunt and Black, 1996] Hunt, A. and Black, A.W. *Unit selection in a concatenative speech synthesis system using a large speech database*. In: IEEE Proceedings of the 1997 International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Atlanta, Georgia, pp. 373-376, 1996.

[Imai, 1983] Imai, S. *Cepstral analysis/synthesis on the Mel frequency scale*. In proc ICASSP83, pp 93-96, Boston, MA. 1983.

[IPA, 1993] IPA: *The International Phonetic Association (revised to 1993) - IPA Chart*, Journal of the International Phonetic Association 23, 1993.

[Jin and Schultz, 2004] Jin, Q. and Schultz, T., "Speaker Segmentation and Clustering in Meetings", *Proceedings of the International Conference of Spoken Language Processing.* Jeju-Island, South Korea, September 2004.

[Latorre et al., 2005] Latorre, J., Iwano, K. and Furui, S. *Polyglot synthesis using a mixture of monolingual corpora*. ICASSP, Philadelpha, PA, 2005.

[Schultz, 2001] Schultz, T. and Waibel, A., "Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition", *Speech Communication, Volume 35, Issue 1-2, pp 31-51, August 2001*.

[Schultz, 2002] Schultz, T., "GlobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University", *Proceedings of the International Conference of Spoken Language Processing.* Denver, CO, September 2002.

[Toda et al., 2004] Toda, T., Black. A. and Tokuda, K. *Mapping from Articulatory Movements to Vocal Tract Spectrum with Gaussian Mixture Models for Articulatory Speech Synthesis*. pp 31-36 5th ISCA Speech Synthesis Workshop, Pittsburgh, PA, 2004.

[Tokuda et al, 2000] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kiramura, T. Speech parameter generation algorithms for HMM-based speech synthesis. In proc. ICASSP2000, pp 1315-1318, Istanbul, Turkey. 2000.

[Tokuda et al., 2002] Tokuda, K., Zen, H. and Black, A. *An HMM-based Speech Synthesis Systems applied to English.* IEEE TTS Workshop, Santa Monica, CA. 2002.